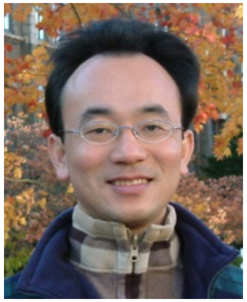
大數據分析
(Big Data Analysis)

大數據分析介紹
(Introduction to Big Data Analysis)

1091BDA01
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
國立臺北大學 資訊管理研究所
https://web.ntpu.edu.tw/~myday

2020-09-16

戴敏育 博士
(Min-Yuh Day, Ph.D.)

國立台北大學 資訊管理研究所 副教授

中央研究院 資訊科學研究所 訪問學人

國立台灣大學 資訊管理 博士

Publications Co-Chairs, IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM 2013-)

Program Co-Chair, IEEE International Workshop on Empirical Methods for Recognizing Inference in TExt (IEEE EM-RITE 2012-)

Publications Chair, The IEEE International Conference on Information Reuse and Integration (IEEE IRI)

aws academy

Accredited Educator

aws certified

Solutions Architect

Associate

aws certified

Cloud Practitioner

大數據分析
(Big Data Analysis)

NTPU

國立臺北大學

National Taipei University

Contact Information

戴敏育 博士 (Min-Yuh Day, Ph.D.)

副教授 (Associate Professor)

國立臺北大學 資訊管理研究所

Institute of Information Management, National Taipei University

電話： 02-86741111 ext. 66873

研究室： 商8F12

地址： 23741 新北市三峽區大學路 151 號

Email：myday@gm.ntpu.edu.tw

網址：http://web.ntpu.edu.tw/~myday/

國立臺北大學
109學年度第1學期
課程大綱
Fall 2020 (2020.09 - 2021.01)

- 課程名稱：**大數據分析 (Big Data Analysis)**
- 授課教師：戴敏育 (Min-Yuh Day)
- 開課系所：資管所碩士班
- 開課資料：選修 半學年 3 學分 (3 Credits, Elective)
- 上課時間：週三 7, 8, 9 (15:10-18:00)
- 上課教室：商8F40 (台北大學三峽校區)

教學目標

1. 瞭解大數據分析基本概念與研究議題。。
2. 具備大數據分析實務操作能力。
3. 進行大數據分析相關之資訊管理研究。

Course Objectives

1. Understand the fundamental concepts and research issues of big data analysis.
2. Equip with Hands-on practices of big data analysis.
3. Conduct information systems research in the context of big data analysis.

內容綱要

- 本課程介紹大數據分析基本概念、研究議題、與實務操作。
- 課程內容包括
  1. 大數據分析介紹
  2. AI人工智慧與大數據分析
  3. Python 大數據分析基礎
  4. Python Pandas 大數據量化分析
  5. Python Scikit-Learn 機器學習
  6. TensorFlow 深度學習金融大數據分析
  7. AI 機器人理財顧問
  8. 金融科技智慧型交談機器人
  9. 金融科技數位沙盒實作
  10. 大數據分析個案研究

Course Outline

- This course introduces the fundamental concepts, research issues, and hands-on practices of big data analysis.
- Topics include
  1. Introduction to Big Data Analysis
  2. AI and Big Data Analysis
  3. Foundations of Big Data Analysis in Python
  4. Quantitative Big Data Analysis with Pandas in Python
  5. Machine Learning with Scikit-Learn In Python
  6. Deep Learning for Finance Big Data Analysis with TensorFlow
  7. Artificial Intelligence for Robo-Advisors
  8. Conversational Commerce and Intelligent Chatbots for Fintech
  9. Hands-on Practices with FintechSpace Digital Sandbox
  10. Case Study on Big Data Analysis

資訊管理研究所
系核心能力
(Core Competence)

- 資訊科技新知探索與系統開發應用 80 %
- 網路行銷企劃能力 10 %
- 論文寫作與獨立研究能力 10 %

校四大基本素養
(Four Fundamental Qualities)

- 專業 (Professionalism)
  - 創意思考與問題解決 (Creative thinking and Problem-solving) 30 %
  - 綜合統整(Comprehensive Integration) 30 %
- 人際 (Interpersonal Relationship)
  - 溝通協調 (Communication and Coordination) 10 %
  - 團隊合作 (Teamwork) 10 %
- 倫理 (Ethics)
  - 誠信正直(Honesty and Integrity) 5 %
  - 尊重自省(Self-Esteem and Self-reflection) 5 %
- 國際觀 (International Vision)
  - 多元關懷 (Caring for Diversity) 5 %
  - 跨界宏觀 (Interdisciplinary Vision) 5 %

商學院學習目標
(College Learning Goals)

- Ethics/Corporate Social Responsibility
- Global Knowledge/Awareness
- Communication
- Analytical and Critical Thinking

系所學習目標
(Department Learning Goals)

- Information Technologies and System Development Capabilities
- Internet Marketing Management Capabilities
- Research capabilities

課程大綱 (Syllabus)

週次 (Week)	日期 (Date)	內容 (Subject/Topics)
1	2020/09/16	大數據分析介紹 (Introduction to Big Data Analysis)
2	2020/09/23	AI人工智慧與大數據分析 (AI and Big Data Analysis)
3	2020/09/30	Python 大數據分析基礎 (Foundations of Big Data Analysis in Python)
4	2020/10/07	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)
5	2020/10/14	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)
6	2020/10/21	Python Pandas 大數據量化分析 (Quantitative Big Data Analysis with Pandas in Python)

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 I)

教學方法與教學活動
(Teaching methods and activities)

- 講授 (Lecture)
- 討論 (Discussion)
- 實習 (Practicum)

評量方式
(Evaluation Methods)

- 個人報告 (Individual Presentation) 60 %
- 團體報告 (Group Presentation) 10 %
- 個案分析報告 (Case Report) 10 %
- 課堂參與 (Class Participation) 10 %
- 作業 (Assignment) 10 %

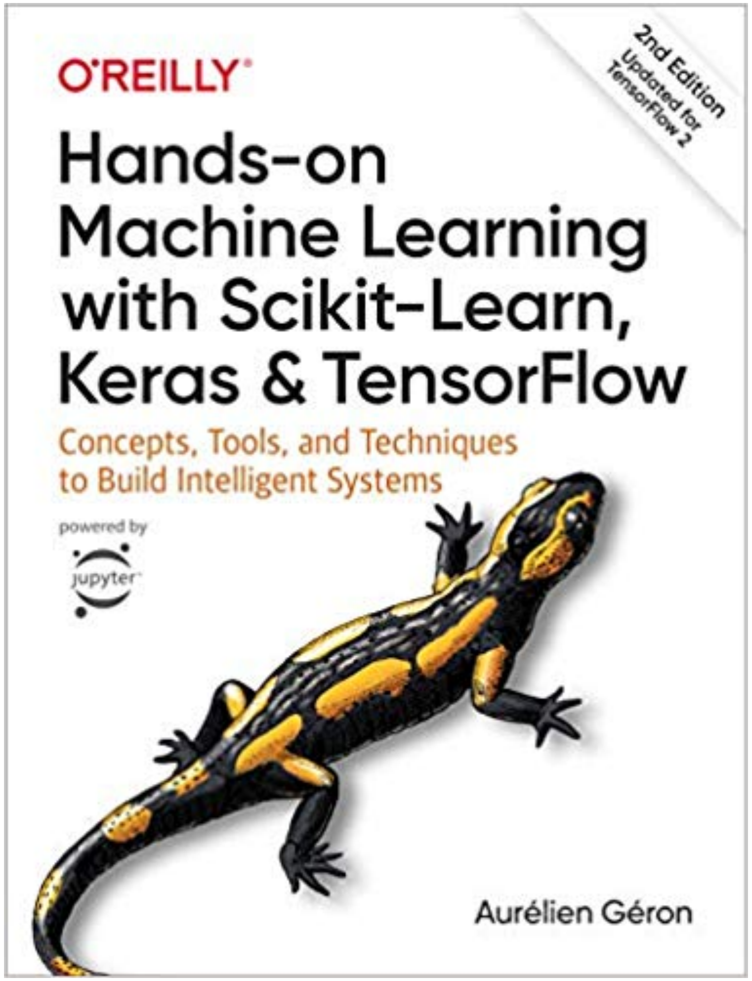
指定用書
(Required Texts)

- Aurélien Géron (2019),
  Hands-On Machine Learning with Scikit-Learn, Keras, and TensorFlow: Concepts, Tools, and Techniques to Build Intelligent Systems,
  2nd Edition, O’Reilly Media.

參考書目
(Reference Books)

- Yves Hilpisch (2018),
  Python for Finance: Mastering Data-Driven Finance,
  2nd Edition, O'Reilly Media.
- 其他參考資料(Other References)：
  - Paolo Sironi (2016), FinTech Innovation: From Robo-Advisors to Goal Based Investing and Gamification, Wiley.
  - Yuxing Yan (2017), Python for Finance: Apply powerful finance models and quantitative analysis with Python, Second Edition, Packt Publishing

Aurélien Géron (2019),
Hands-On Machine Learning with Scikit-Learn, Keras, and TensorFlow: Concepts, Tools, and Techniques to Build Intelligent Systems, 2nd Edition
O'Reilly Media, 2019


O'REILLY
2nd Edition
Updated for TensorFlow 2
Hands-on Machine Learning with Scikit-Learn, Keras & TensorFlow
Concepts, Tools, and Techniques to Build Intelligent Systems
powered by
jupyter
Aurélien Géron

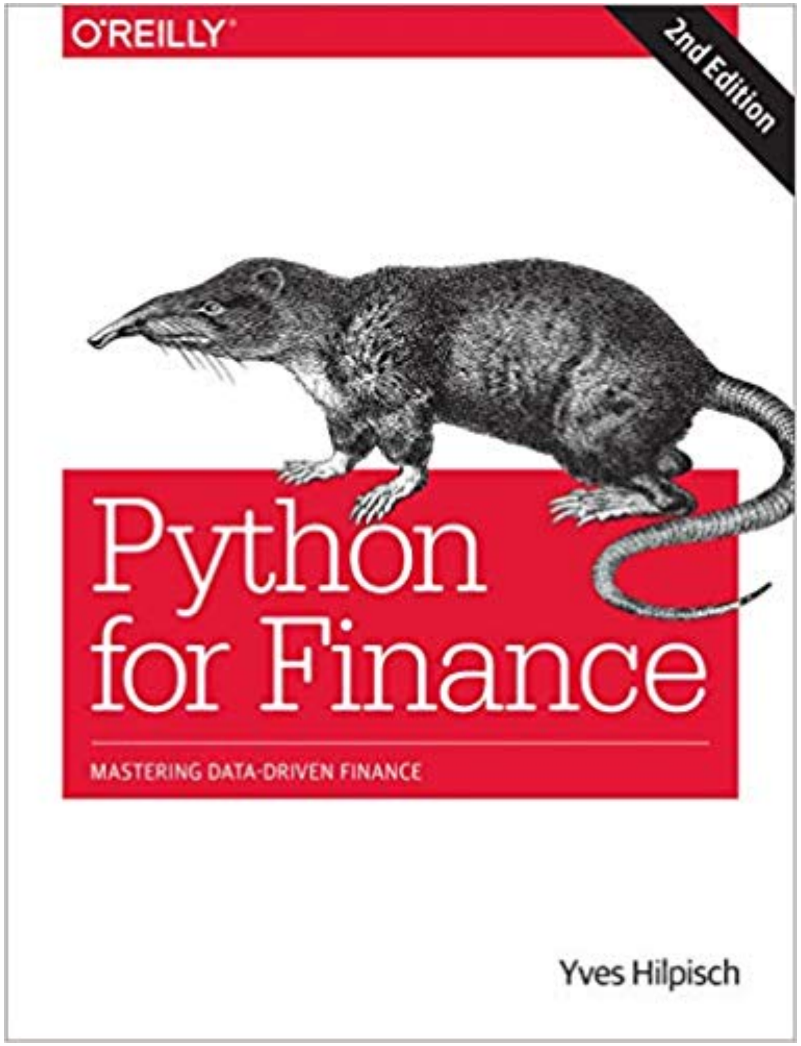
https://github.com/ageron/handson-ml2

Source: https://www.amazon.com/Hands-Machine-Learning-Scikit-Learn-TensorFlow/dp/1492032646/

Yves Hilpisch (2018),

Python for Finance: Mastering Data-Driven Finance,

O'Reilly



Source: https://www.amazon.com/Python-Finance-Mastering-Data-Driven/dp/1492024333

Paolo Sironi (2016)

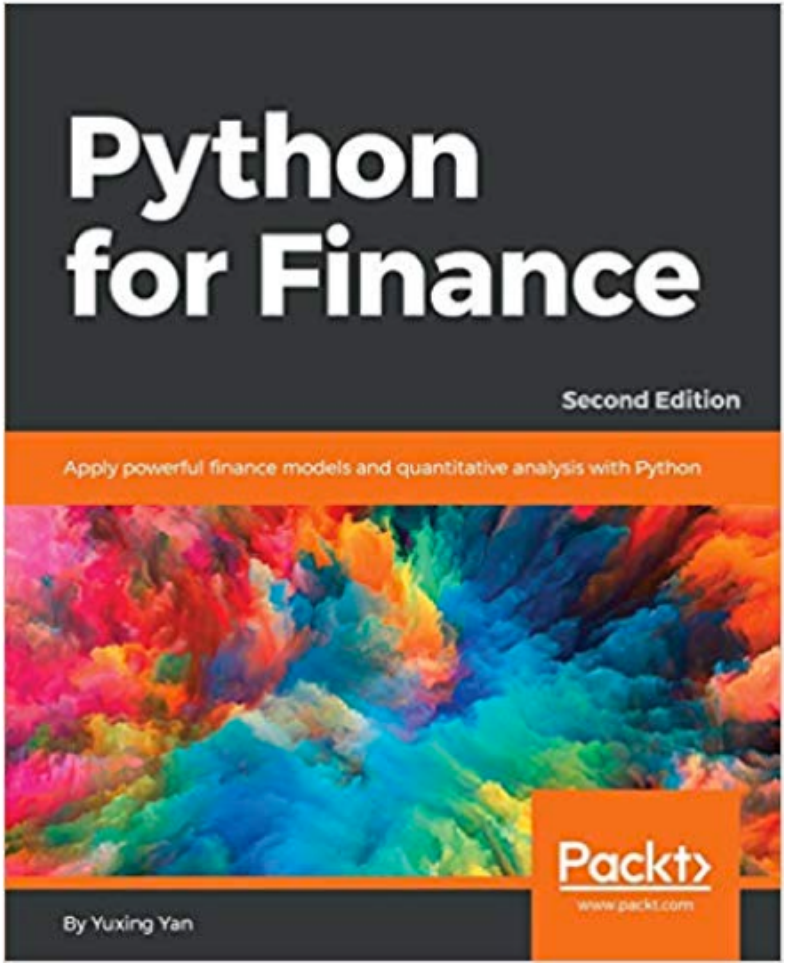
FinTech Innovation:

From Robo-Advisors to Goal Based Investing and Gamification,

Wiley

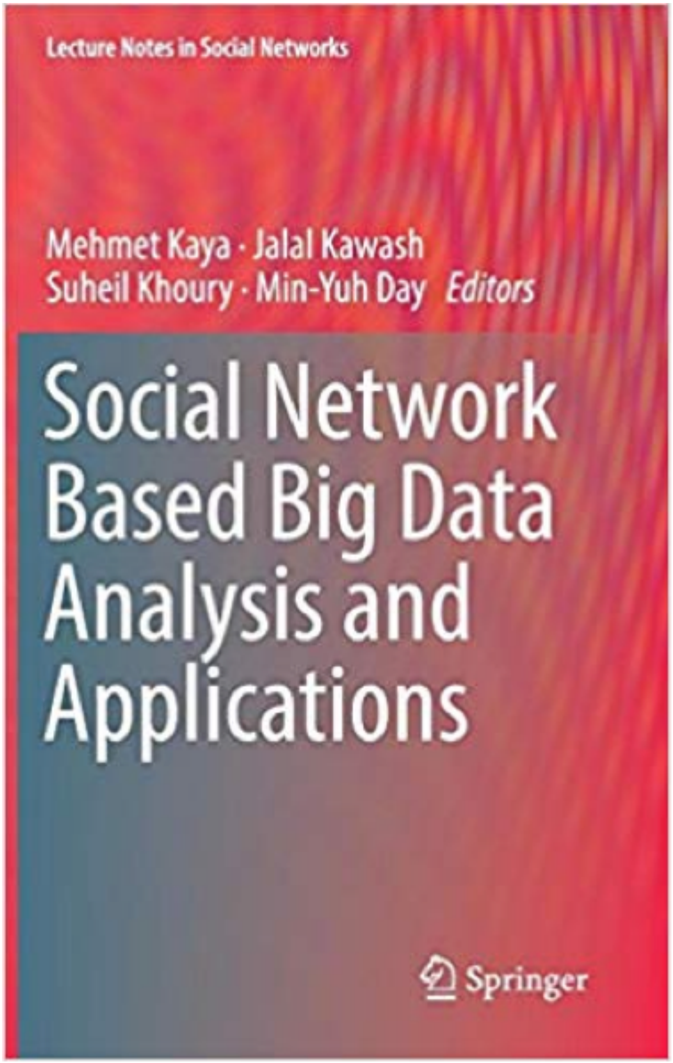
Source: https://www.amazon.com/FinTech-Innovation-Robo-Advisors-Investing-Gamification/dp/1119226988

Yuxing Yan (2017), Python for Finance: Apply powerful finance models and quantitative analysis with Python, Second Edition, Packt Publishing


Python
for Finance
Second Edition
Apply powerful finance models and quantitative analysis with Python
Packt›
www.packt.com
By Yuxing Yan

Source: https://www.amazon.com/Python-Finance-powerful-quantitative-analysis/dp/1787125696

Social Network Based Big Data Analysis and Applications, Lecture Notes in Social Networks, Mehmet Kaya, Jalal Kawash, Suheil Khoury, Min-Yuh Day, Springer International Publishing, 2018.



Lecture Notes in Social Networks

Mehmet Kaya · Jalal Kawash
Suheil Khoury · Min-Yuh Day *Editors*

Social Network Based Big Data Analysis and Applications

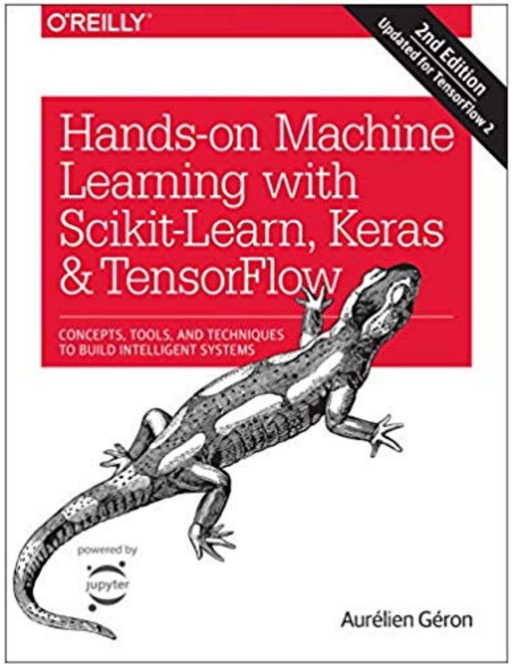
Springer

Source: https://www.amazon.com/Network-Analysis-Applications-Lecture-Networks/dp/3319781952

Hands-On Machine Learning with Scikit-Learn, Keras, and TensorFlow

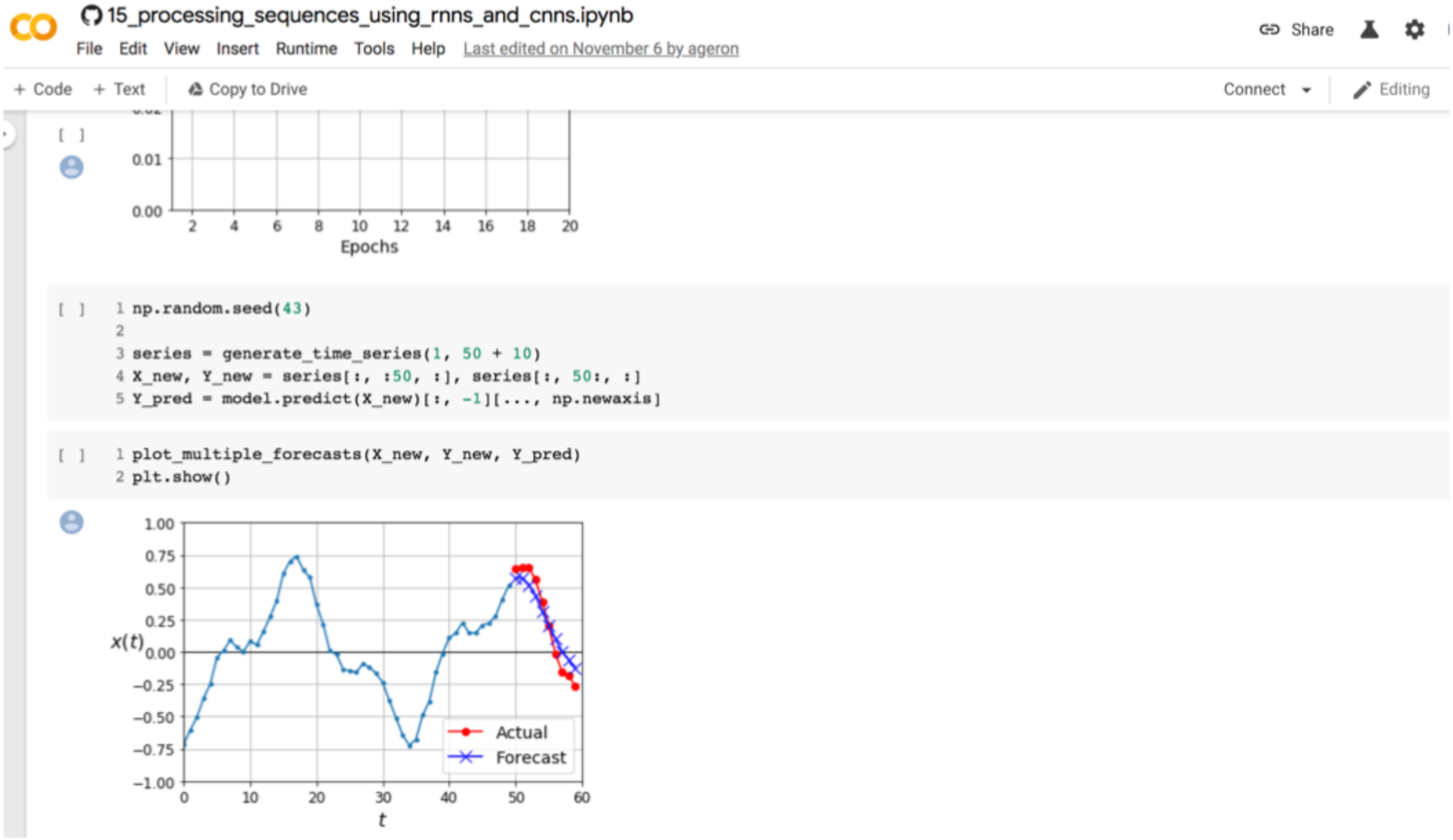
Notebooks

1. The Machine Learning landscape
2. End-to-end Machine Learning project
3. Classification
4. Training Models
5. Support Vector Machines
6. Decision Trees
7. Ensemble Learning and Random Forests
8. Dimensionality Reduction
9. Unsupervised Learning Techniques
10. Artificial Neural Nets with Keras
11. Training Deep Neural Networks
12. Custom Models and Training with TensorFlow
13. Loading and Preprocessing Data
14. Deep Computer Vision Using Convolutional Neural Networks
15. Processing Sequences Using RNNs and CNNs
16. Natural Language Processing with RNNs and Attention
17. Representation Learning Using Autoencoders
18. Reinforcement Learning
19. Training and Deploying TensorFlow Models at Scale


O'REILLY
2nd Edition
Updated for TensorFlow 2
Hands-on Machine Learning with Scikit-Learn, Keras & TensorFlow
CONCEPTS, TOOLS, AND TECHNIQUES TO BUILD INTELLIGENT SYSTEMS
powered by
jupyter
Aurélien Géron

https://github.com/ageron/handson-ml2

Sequences using RNNs and CNNs

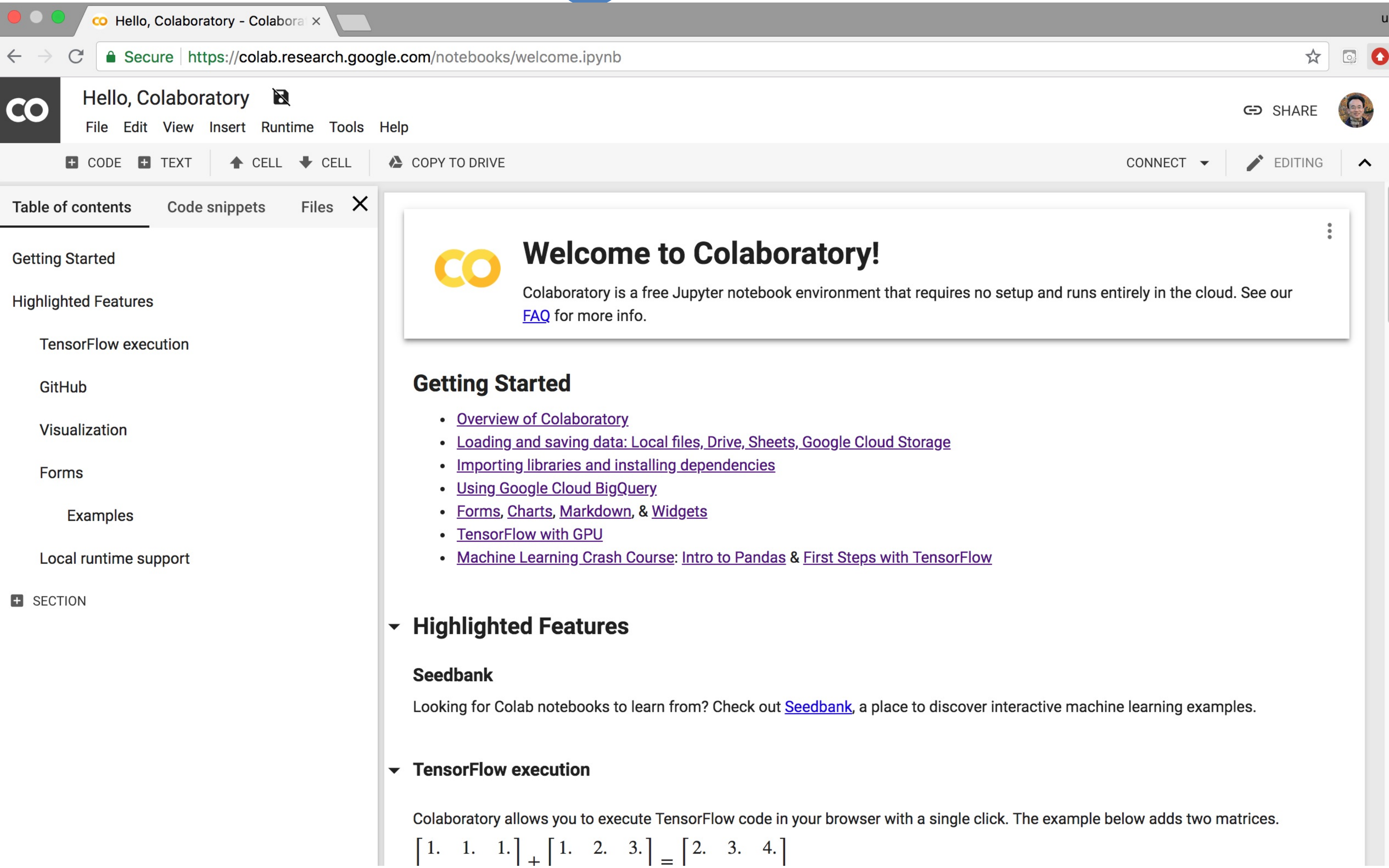


```
np.random.seed(43)

series = generate_time_series(1, 50 + 10)
X_new, Y_new = series[:, :50, :], series[:, 50:, :]
Y_pred = model.predict(X_new)[:, -1][..., np.newaxis]
```

```
plot_multiple_forecasts(X_new, Y_new, Y_pred)
plt.show()
```

https://colab.research.google.com/github/ageron/handson-ml2/blob/master/15_processing_sequences_using_rnns_and_cnns.ipynb

Google Colab



Welcome to Colaboratory!

Colaboratory is a free Jupyter notebook environment that requires no setup and runs entirely in the cloud. See our FAQ for more info.

Getting Started

- Overview of Colaboratory
- Loading and saving data: Local files, Drive, Sheets, Google Cloud Storage
- Importing libraries and installing dependencies
- Using Google Cloud BigQuery
- Forms, Charts, Markdown, & Widgets
- TensorFlow with GPU
- Machine Learning Crash Course: Intro to Pandas & First Steps with TensorFlow

Highlighted Features

Seedbank

Looking for Colab notebooks to learn from? Check out Seedbank, a place to discover interactive machine learning examples.

TensorFlow execution

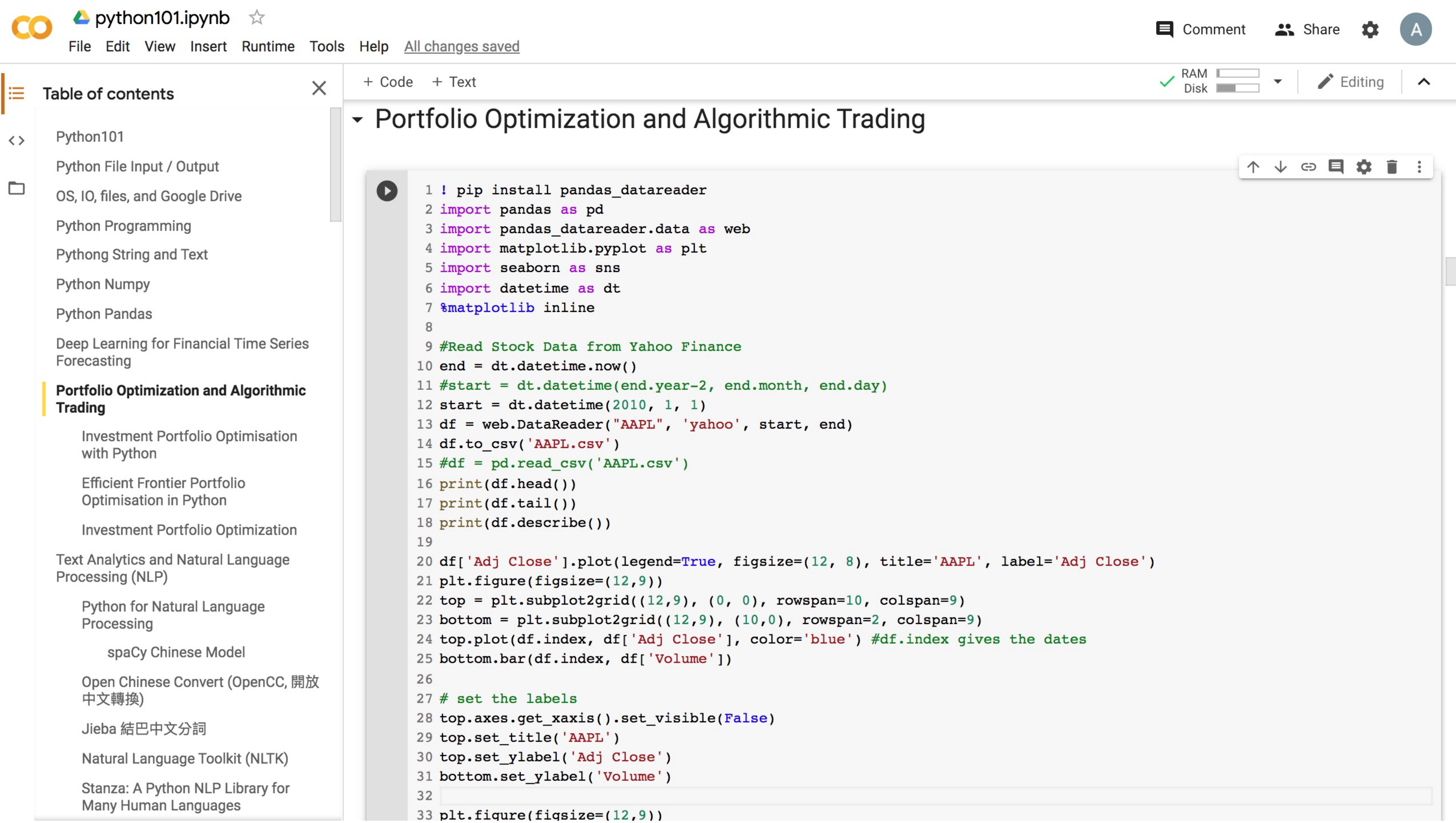
Colaboratory allows you to execute TensorFlow code in your browser with a single click. The example below adds two matrices.

$$\begin{bmatrix} 1. & 1. & 1. \end{bmatrix} + \begin{bmatrix} 1. & 2. & 3. \end{bmatrix} = \begin{bmatrix} 2. & 3. & 4. \end{bmatrix}$$

https://colab.research.google.com/notebooks/welcome.ipynb

Python in Google Colab (Python101)

https://colab.research.google.com/drive/1FEG6DnGvwfUbeo4zJ1zTunjMqf2RkCrT



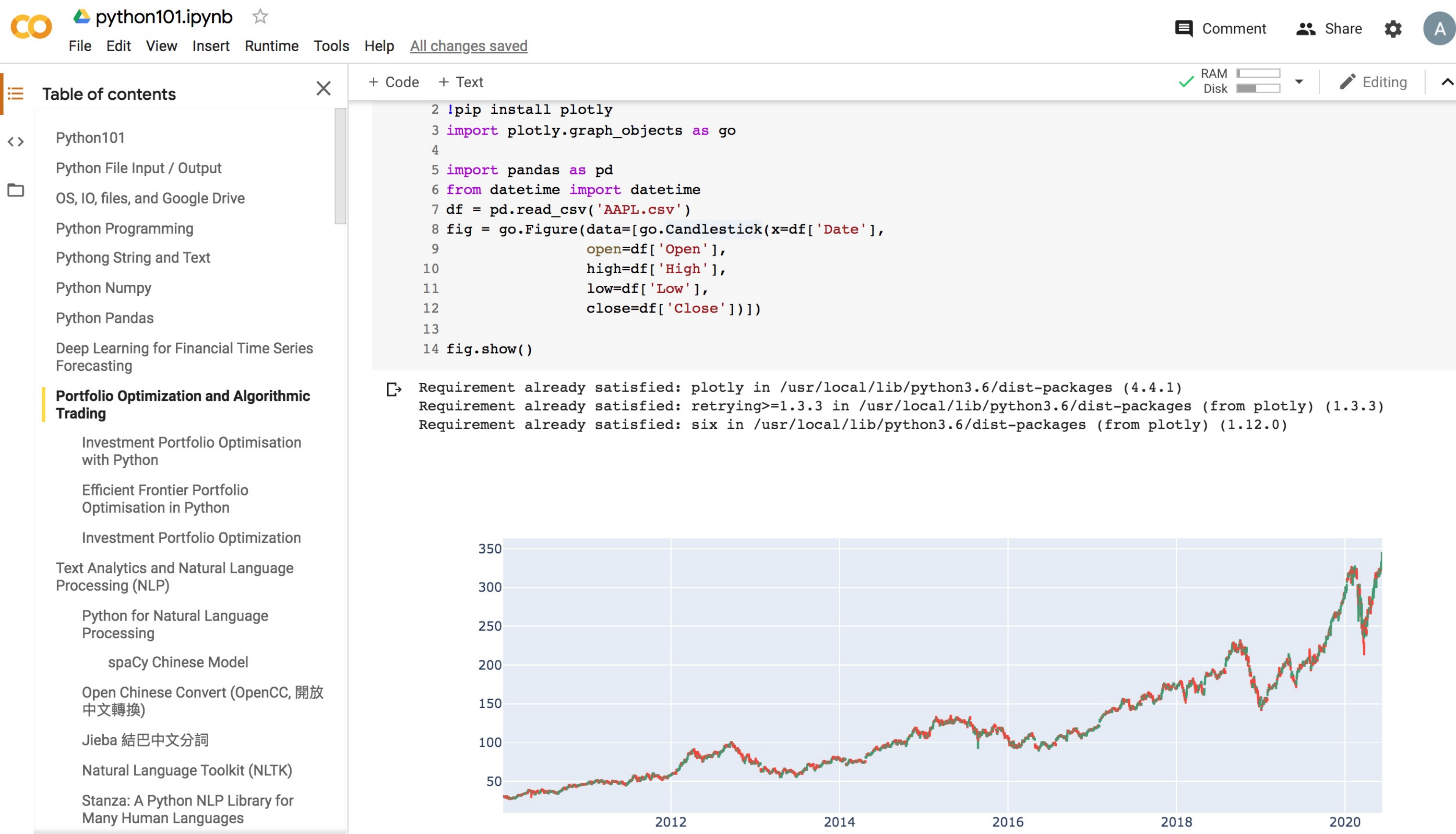
Portfolio Optimization and Algorithmic Trading

```
! pip install pandas_datareader
import pandas as pd
import pandas_datareader.data as web
import matplotlib.pyplot as plt
import seaborn as sns
import datetime as dt
%matplotlib inline

#Read Stock Data from Yahoo Finance
end = dt.datetime.now()
#start = dt.datetime(end.year-2, end.month, end.day)
start = dt.datetime(2010, 1, 1)
df = web.DataReader("AAPL", 'yahoo', start, end)
df.to_csv('AAPL.csv')
#df = pd.read_csv('AAPL.csv')
print(df.head())
print(df.tail())
print(df.describe())

df['Adj Close'].plot(legend=True, figsize=(12, 8), title='AAPL', label='Adj Close')
plt.figure(figsize=(12,9))
top = plt.subplot2grid((12,9), (0, 0), rowspan=10, colspan=9)
bottom = plt.subplot2grid((12,9), (10,0), rowspan=2, colspan=9)
top.plot(df.index, df['Adj Close'], color='blue') #df.index gives the dates
bottom.bar(df.index, df['Volume'])

# set the labels
top.axes.get_xaxis().set_visible(False)
top.set_title('AAPL')
top.set_ylabel('Adj Close')
bottom.set_ylabel('Volume')

plt.figure(figsize=(12,9))
```

http://tinyurl.com/aintpupython101

Python in Google Colab (Python101)

https://colab.research.google.com/drive/1FEG6DnGvwfUbeo4zJ1zTunjMqf2RkCrT



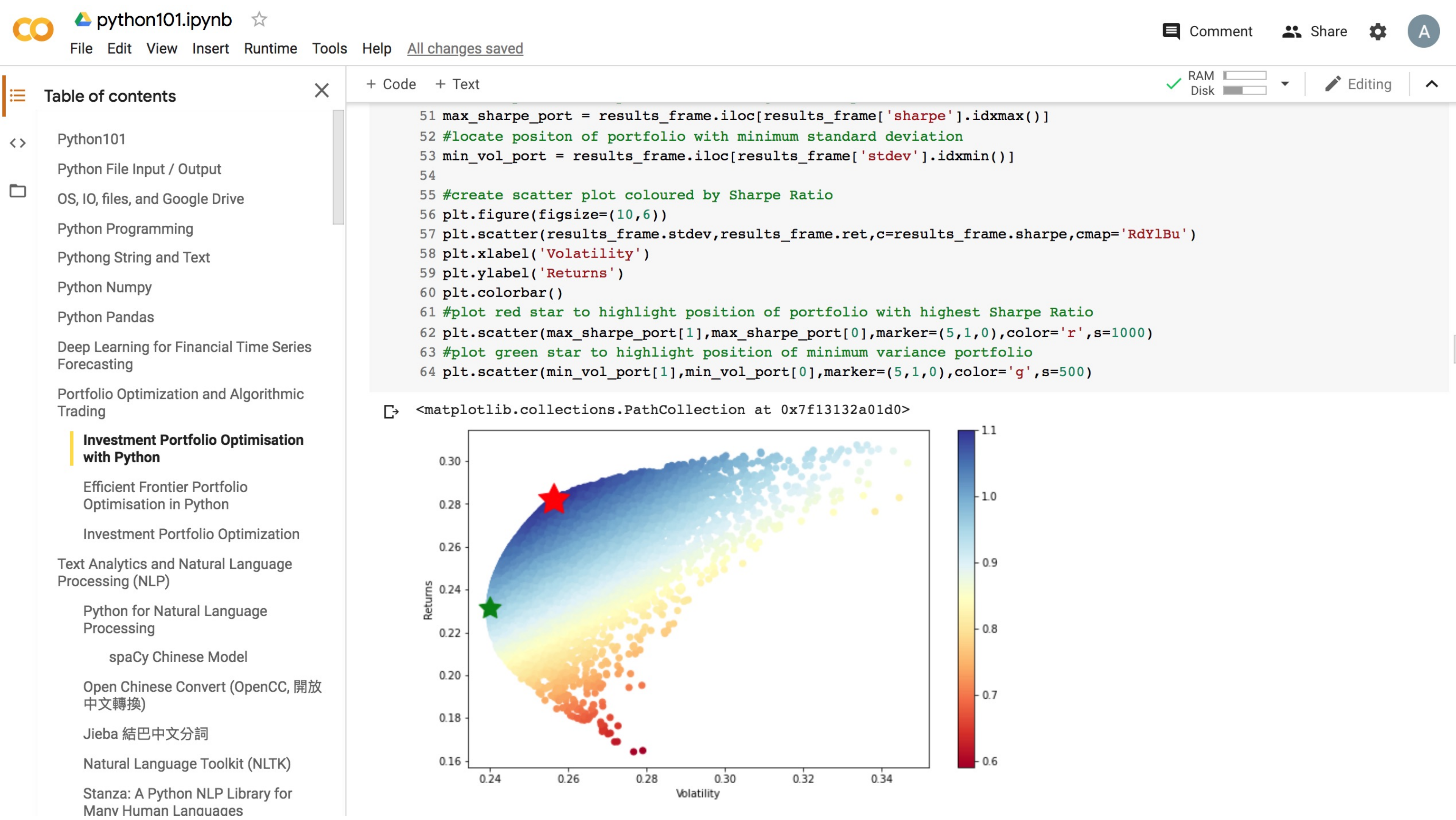
```
!pip install plotly
import plotly.graph_objects as go

import pandas as pd
from datetime import datetime
df = pd.read_csv('AAPL.csv')
fig = go.Figure(data=[go.Candlestick(x=df['Date'],
                open=df['Open'],
                high=df['High'],
                low=df['Low'],
                close=df['Close'])])

fig.show()
```

```
Requirement already satisfied: plotly in /usr/local/lib/python3.6/dist-packages (4.4.1)
Requirement already satisfied: retrying>=1.3.3 in /usr/local/lib/python3.6/dist-packages (from plotly) (1.3.3)
Requirement already satisfied: six in /usr/local/lib/python3.6/dist-packages (from plotly) (1.12.0)
```

http://tinyurl.com/aintpupython101

Python in Google Colab (Python101)

https://colab.research.google.com/drive/1FEG6DnGvwfUbeo4zJ1zTunjMqf2RkCrT



- Python101
- Python File Input / Output
- OS, IO, files, and Google Drive
- Python Programming
- Pythong String and Text
- Python Numpy
- Python Pandas
- Deep Learning for Financial Time Series Forecasting
- Portfolio Optimization and Algorithmic Trading
  - **Investment Portfolio Optimisation with Python**
  - Efficient Frontier Portfolio Optimisation in Python
  - Investment Portfolio Optimization
- Text Analytics and Natural Language Processing (NLP)
  - Python for Natural Language Processing
    - spaCy Chinese Model
  - Open Chinese Convert (OpenCC, 開放中文轉換)
  - Jieba 結巴中文分詞
  - Natural Language Toolkit (NLTK)
  - Stanza: A Python NLP Library for Many Human Languages

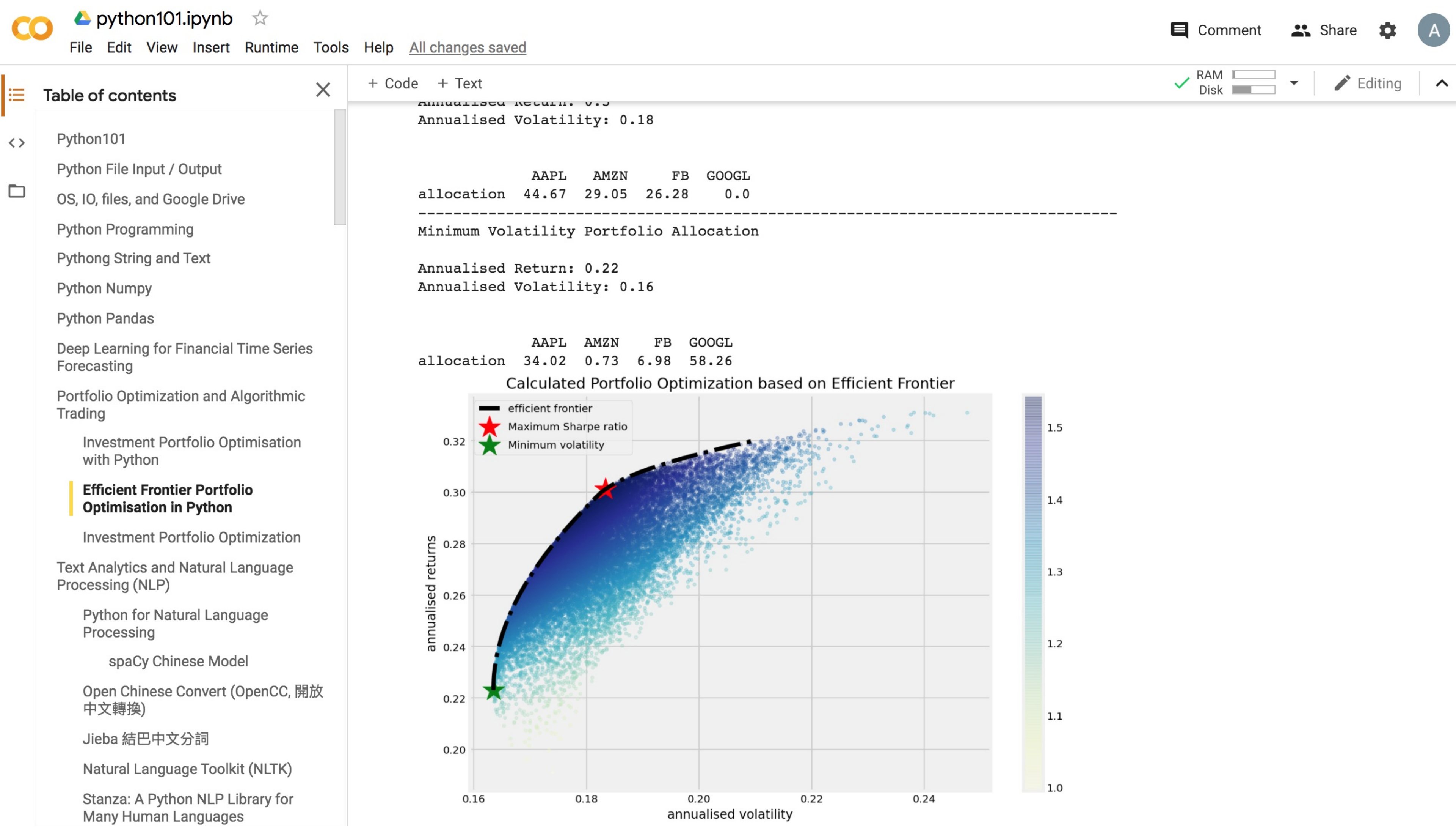
```
max_sharpe_port = results_frame.iloc[results_frame['sharpe'].idxmax()]
#locate positon of portfolio with minimum standard deviation
min_vol_port = results_frame.iloc[results_frame['stdev'].idxmin()]

#create scatter plot coloured by Sharpe Ratio
plt.figure(figsize=(10,6))
plt.scatter(results_frame.stdev,results_frame.ret,c=results_frame.sharpe,cmap='RdYlBu')
plt.xlabel('Volatility')
plt.ylabel('Returns')
plt.colorbar()
#plot red star to highlight position of portfolio with highest Sharpe Ratio
plt.scatter(max_sharpe_port[1],max_sharpe_port[0],marker=(5,1,0),color='r',s=1000)
#plot green star to highlight position of minimum variance portfolio
plt.scatter(min_vol_port[1],min_vol_port[0],marker=(5,1,0),color='g',s=500)
```

```
<matplotlib.collections.PathCollection at 0x7f13132a01d0>
```

http://tinyurl.com/aintpupython101

Python in Google Colab (Python101)

https://colab.research.google.com/drive/1FEG6DnGvwfUbeo4zJ1zTunjMqf2RkCrT



Python101

Python File Input / Output

OS, IO, files, and Google Drive

Python Programming

Pythong String and Text

Python Numpy

Python Pandas

Deep Learning for Financial Time Series Forecasting

Portfolio Optimization and Algorithmic Trading

Investment Portfolio Optimisation with Python

Efficient Frontier Portfolio Optimisation in Python

Investment Portfolio Optimization

Text Analytics and Natural Language Processing (NLP)

Python for Natural Language Processing

spaCy Chinese Model

Open Chinese Convert (OpenCC, 開放中文轉換)

Jieba 結巴中文分詞

Natural Language Toolkit (NLTK)

Stanza: A Python NLP Library for Many Human Languages

```
Annualised Return: 0.3
Annualised Volatility: 0.18


             AAPL   AMZN     FB  GOOGL
allocation  44.67  29.05  26.28    0.0
--------------------------------------------------------------------------------
Minimum Volatility Portfolio Allocation

Annualised Return: 0.22
Annualised Volatility: 0.16


             AAPL  AMZN    FB  GOOGL
allocation  34.02  0.73  6.98  58.26
```

Calculated Portfolio Optimization based on Efficient Frontier

http://tinyurl.com/aintpupython101

Python in Google Colab (Python101)

https://colab.research.google.com/drive/1FEG6DnGvwfUbeo4zJ1zTunjMqf2RkCrT

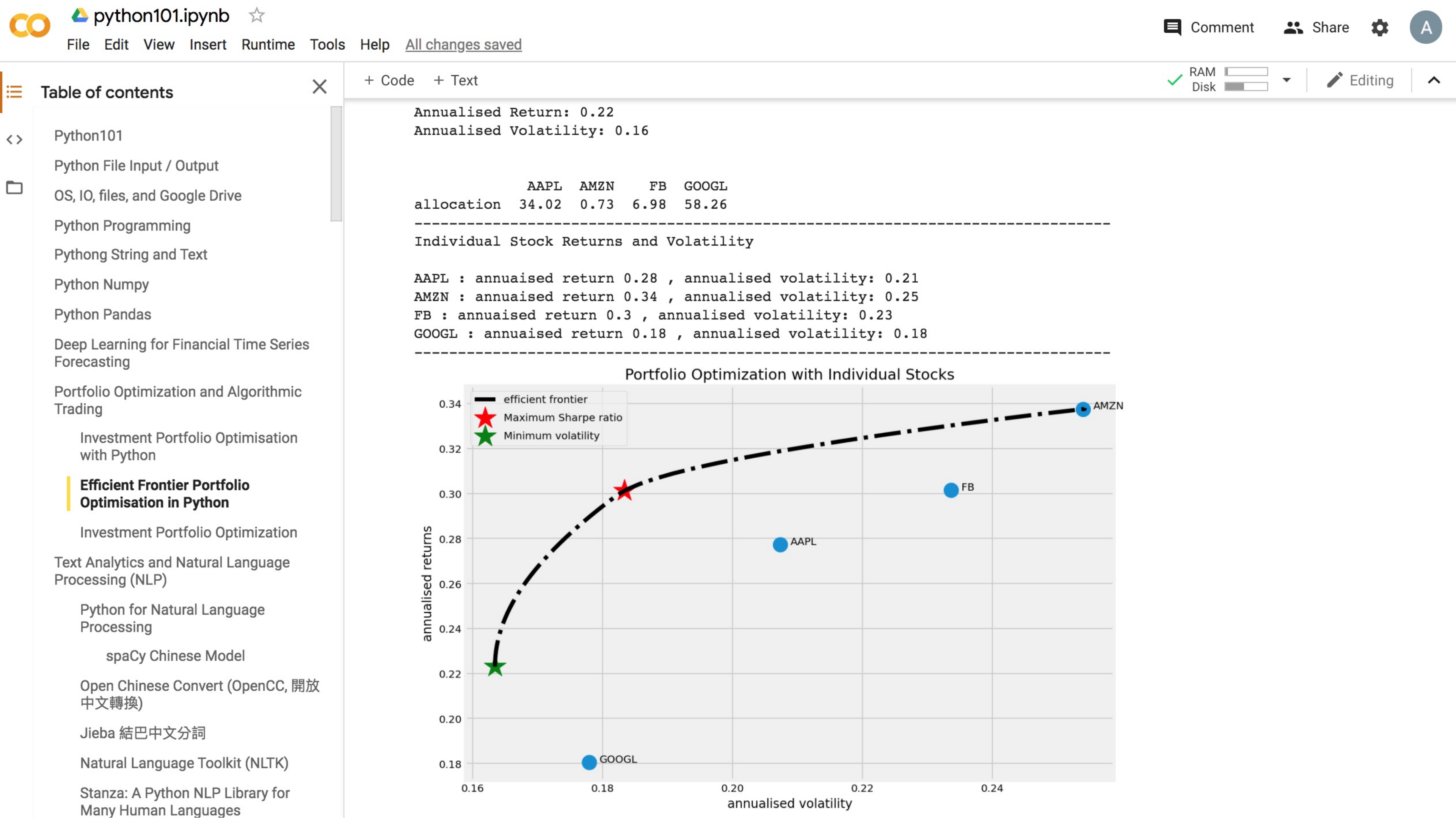


Table of contents

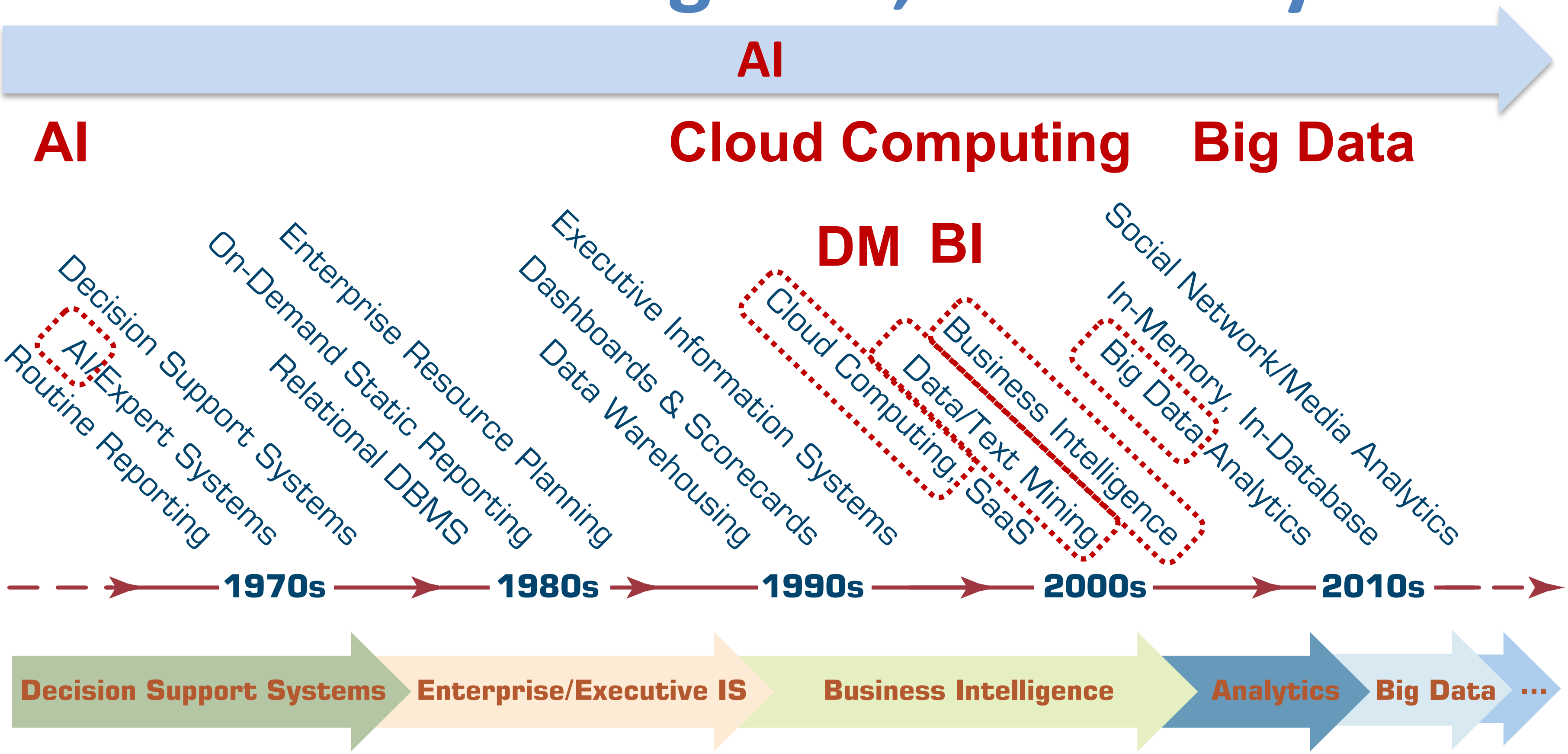
```
Annualised Return: 0.22
Annualised Volatility: 0.16


              AAPL  AMZN    FB  GOOGL
allocation   34.02  0.73  6.98  58.26
--------------------------------------------------------------------------------
Individual Stock Returns and Volatility

AAPL : annuaised return 0.28 , annualised volatility: 0.21
AMZN : annuaised return 0.34 , annualised volatility: 0.25
FB : annuaised return 0.3 , annualised volatility: 0.23
GOOGL : annuaised return 0.18 , annualised volatility: 0.18
--------------------------------------------------------------------------------
```

http://tinyurl.com/aintpupython101

Big Data Analysis

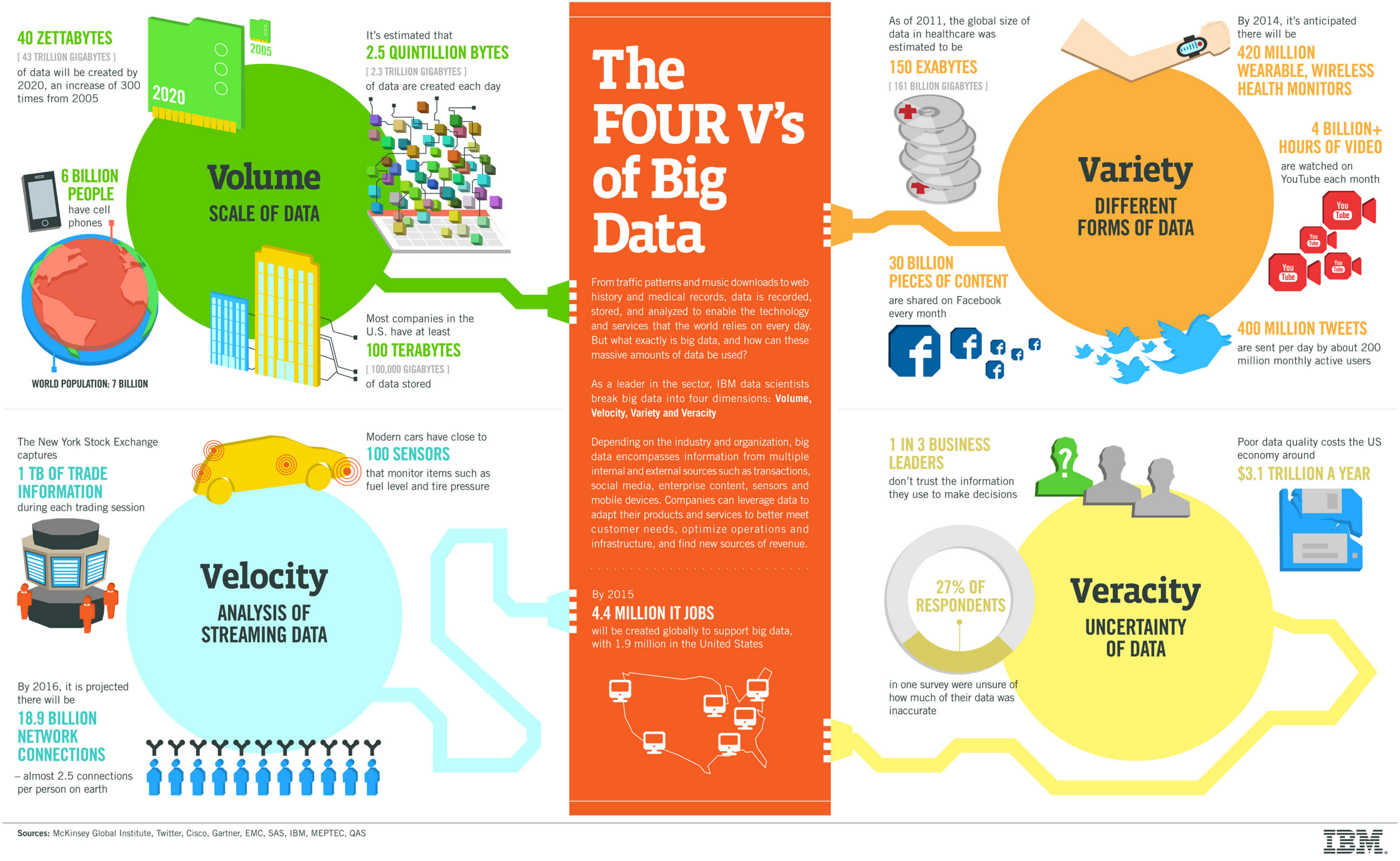
AI, Big Data, Cloud Computing

Evolution of Decision Support, Business Intelligence, and Analytics



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

Big Data 4 V



Volume

SCALE OF DATA

40 ZETTABYTES
[43 TRILLION GIGABYTES]
of data will be created by 2020, an increase of 300 times from 2005

It's estimated that
2.5 QUINTILLION BYTES
[2.3 TRILLION GIGABYTES]
of data are created each day

6 BILLION PEOPLE
have cell phones

WORLD POPULATION: 7 BILLION

Most companies in the U.S. have at least
100 TERABYTES
[100,000 GIGABYTES]
of data stored

Velocity

ANALYSIS OF STREAMING DATA

The New York Stock Exchange captures
1 TB 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

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

Variety

DIFFERENT FORMS OF DATA

As of 2011, the global size of data in healthcare was estimated to be
150 EXABYTES
[161 BILLION 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

UNCERTAINTY OF DATA

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

27% OF RESPONDENTS
in one survey were unsure of how much of their data was inaccurate

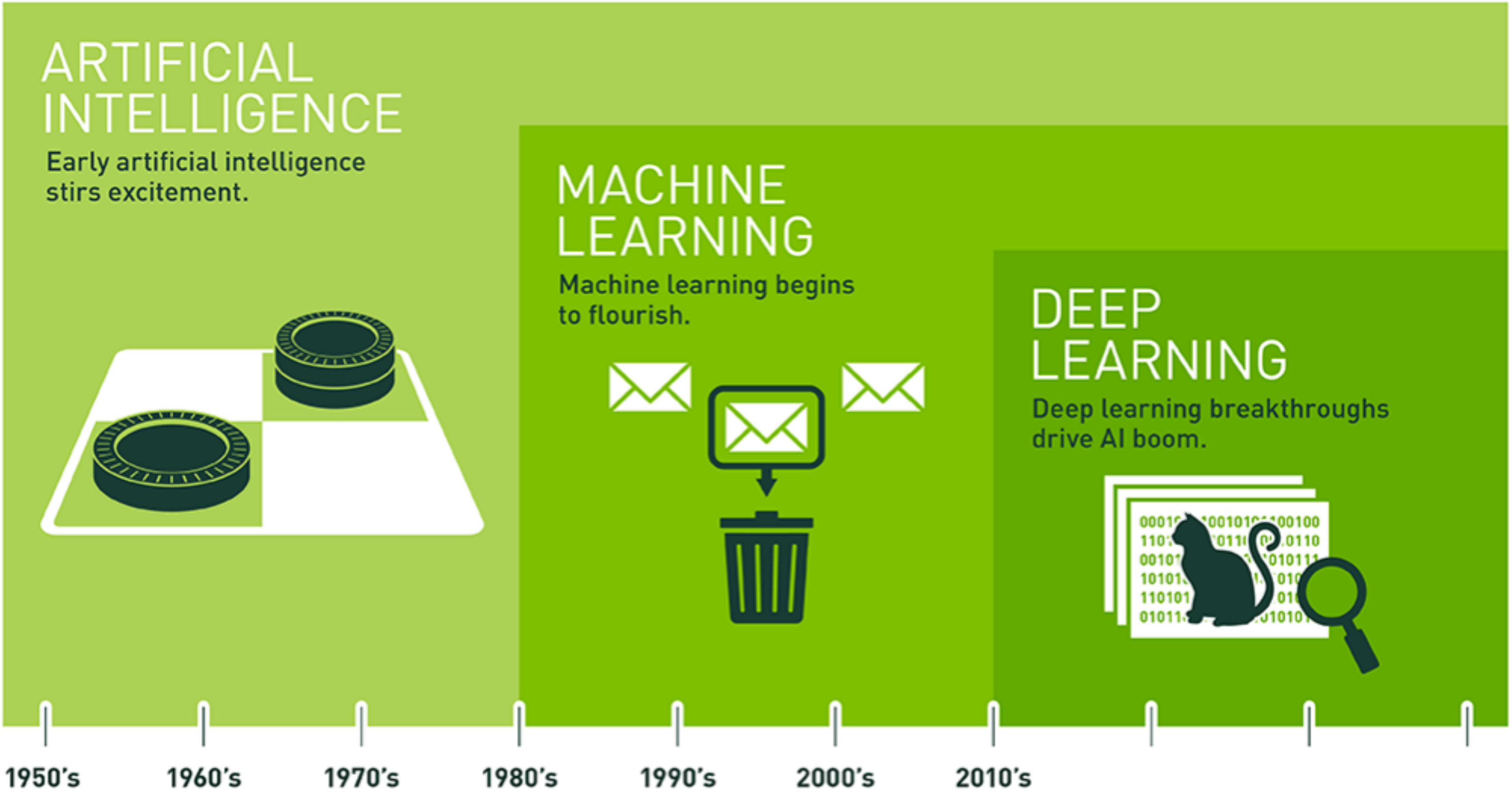
Sources: McKinsey Global Institute, Twitter, Cisco, Gartner, EMC, SAS, IBM, MEPTEC, QAS

IBM

Source: https://www-01.ibm.com/software/data/bigdata/

Value

Artificial Intelligence
Machine Learning & Deep Learning



ARTIFICIAL INTELLIGENCE

Early artificial intelligence stirs excitement.

MACHINE LEARNING

Machine learning begins to flourish.

DEEP LEARNING

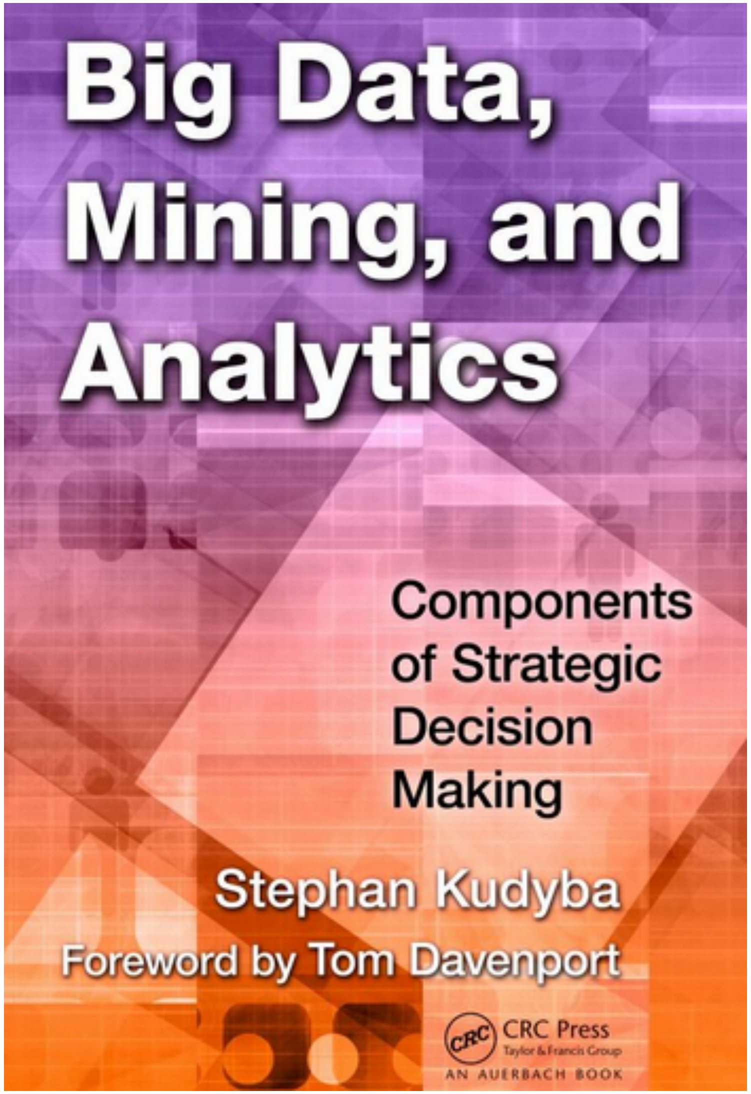
Deep learning breakthroughs drive AI boom.

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.

Source: https://blogs.nvidia.com/blog/2016/07/29/whats-difference-artificial-intelligence-machine-learning-deep-learning-ai/

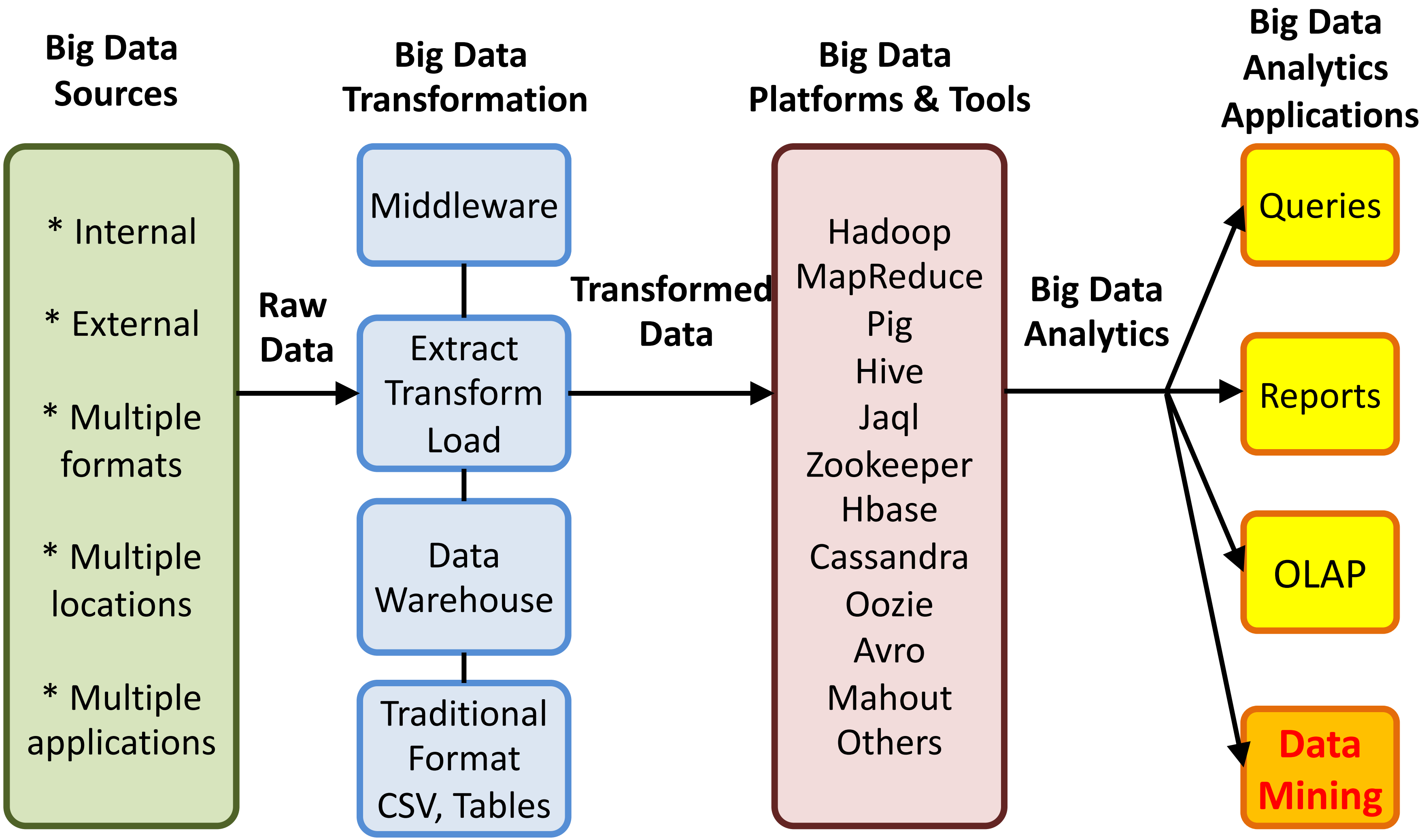
Stephan Kudyba (2014),

Big Data, Mining, and Analytics: Components of Strategic Decision Making, Auerbach Publications


Big Data, Mining, and Analytics
Components of Strategic Decision Making
Stephan Kudyba
Foreword by Tom Davenport
CRC Press
Taylor & Francis Group
AN AUERBACH BOOK

Source: http://www.amazon.com/gp/product/1466568704

Architecture of Big Data Analytics



Big Data Sources

- * Internal
- * External
- * Multiple formats
- * Multiple locations
- * Multiple applications

Raw Data

Big Data Transformation

Middleware

Extract Transform Load

Data Warehouse

Traditional Format CSV, Tables

Transformed Data

Big Data Platforms & Tools

Hadoop
MapReduce
Pig
Hive
Jaql
Zookeeper
Hbase
Cassandra
Oozie
Avro
Mahout
Others

Big Data Analytics

Big Data Analytics Applications

Queries

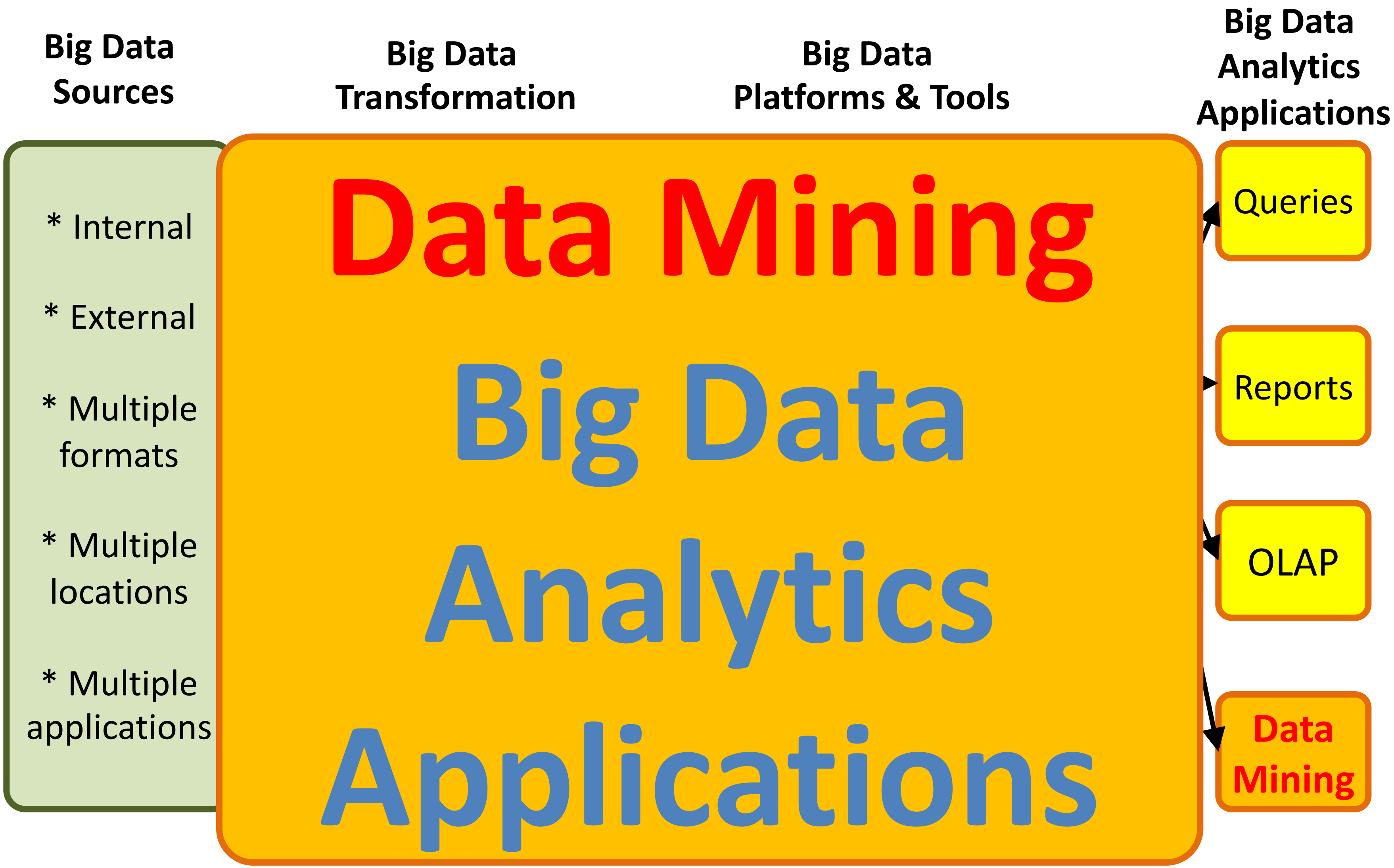
Reports

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

Big Data Transformation

Big Data Platforms & Tools

Big Data Analytics Applications

* Internal

* External

* Multiple formats

* Multiple locations

* Multiple applications

Data Mining

Big Data

Analytics

Applications

Queries

Reports

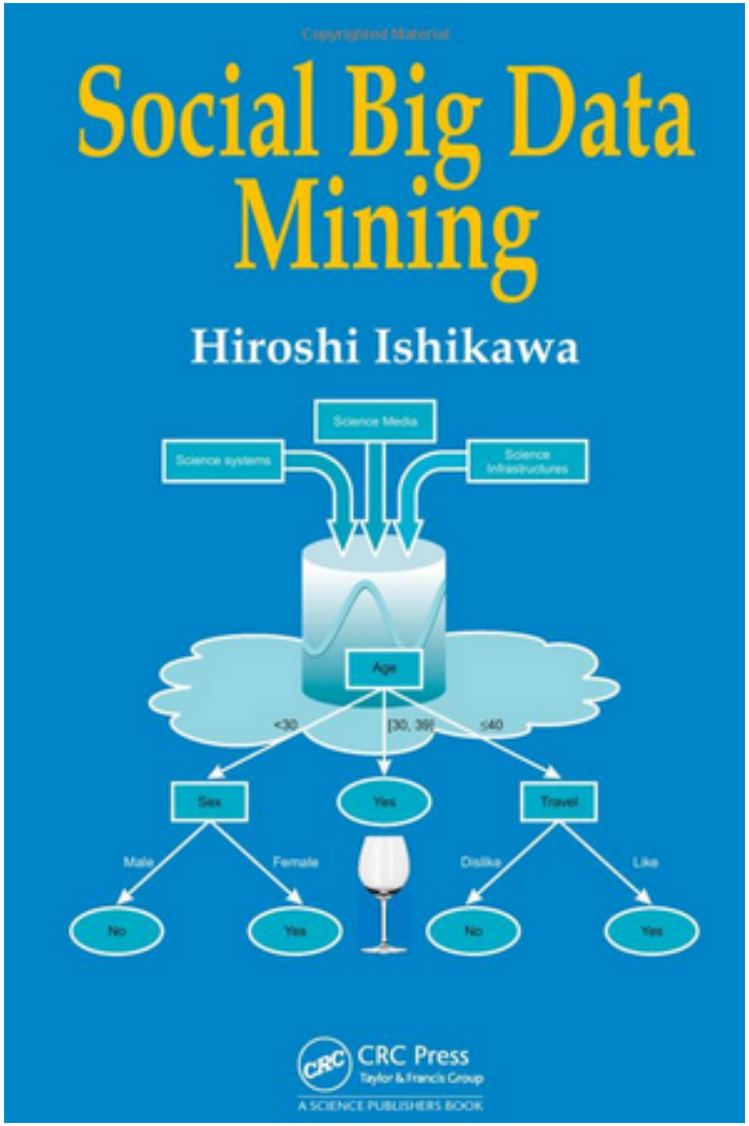
OLAP

Data Mining

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

Social Big Data Mining

(Hiroshi Ishikawa, 2015)



Source: http://www.amazon.com/Social-Data-Mining-Hiroshi-Ishikawa/dp/149871093X

Architecture for Social Big Data Mining

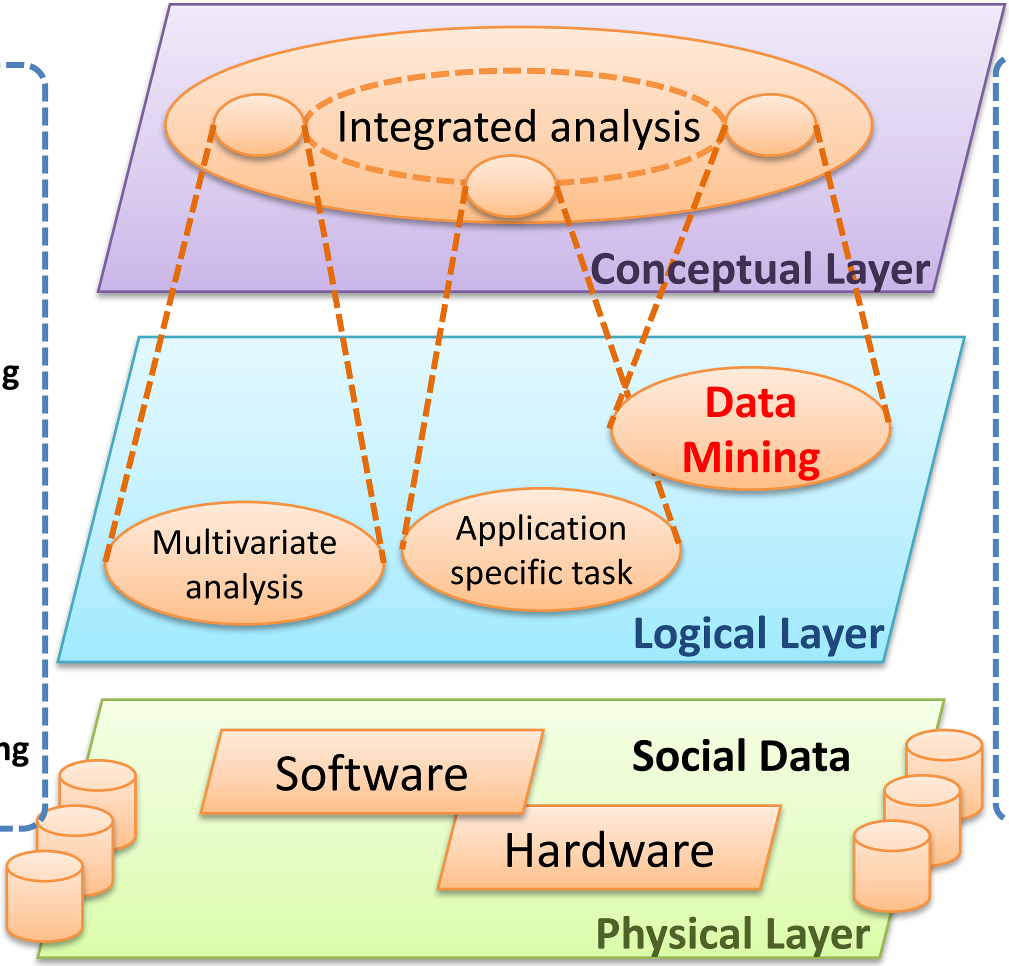
(Hiroshi Ishikawa, 2015)

Enabling Technologies

- Integrated analysis model

- Natural Language Processing
- Information Extraction
- Anomaly Detection
- Discovery of relationships among heterogeneous data
- Large-scale visualization

- Parallel distrusted processing



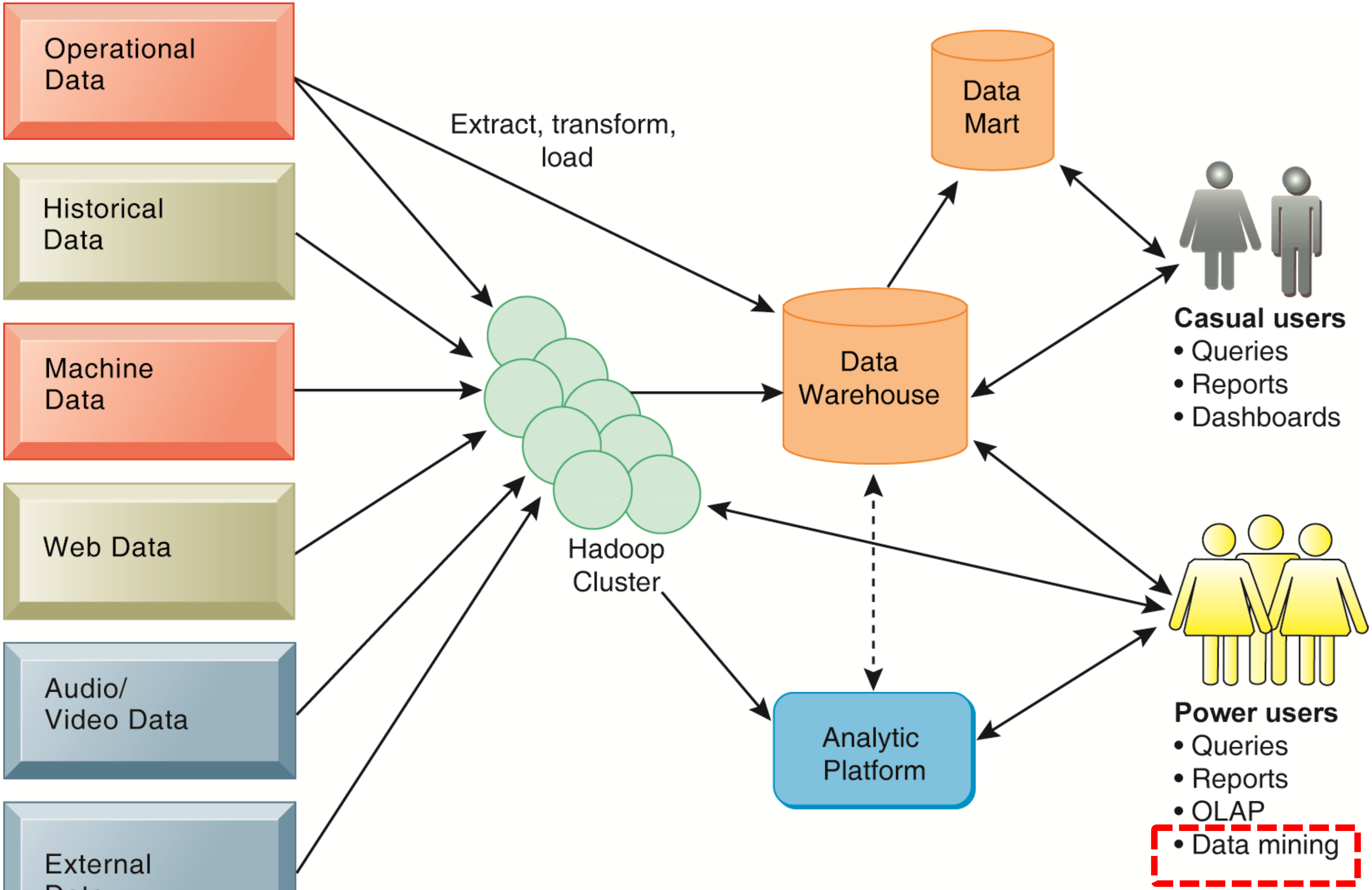
Analysts

- 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

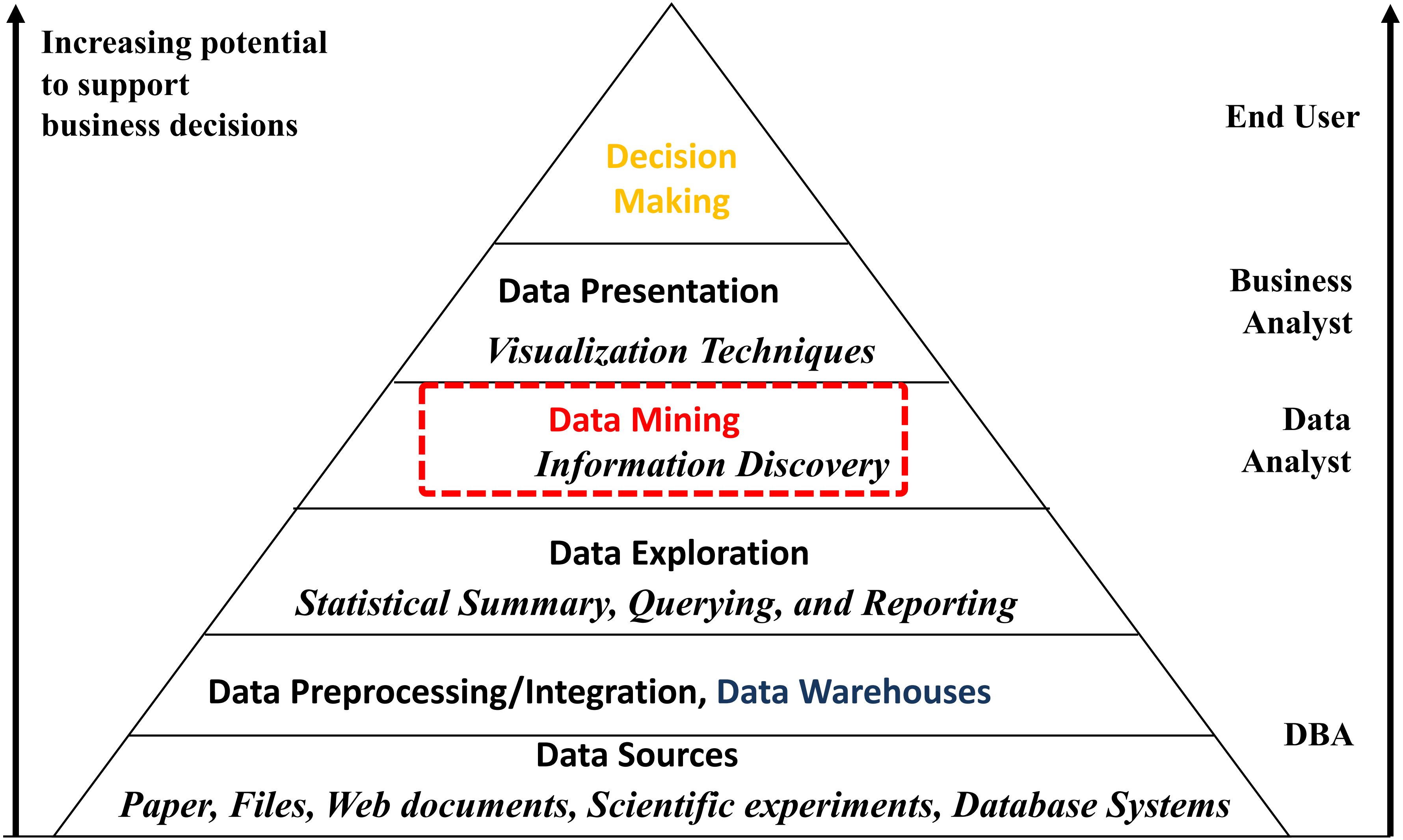
Business Intelligence (BI) Infrastructure



Source: Kenneth C. Laudon & Jane P. Laudon (2014), Management Information Systems: Managing the Digital Firm, Thirteenth Edition, Pearson.

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

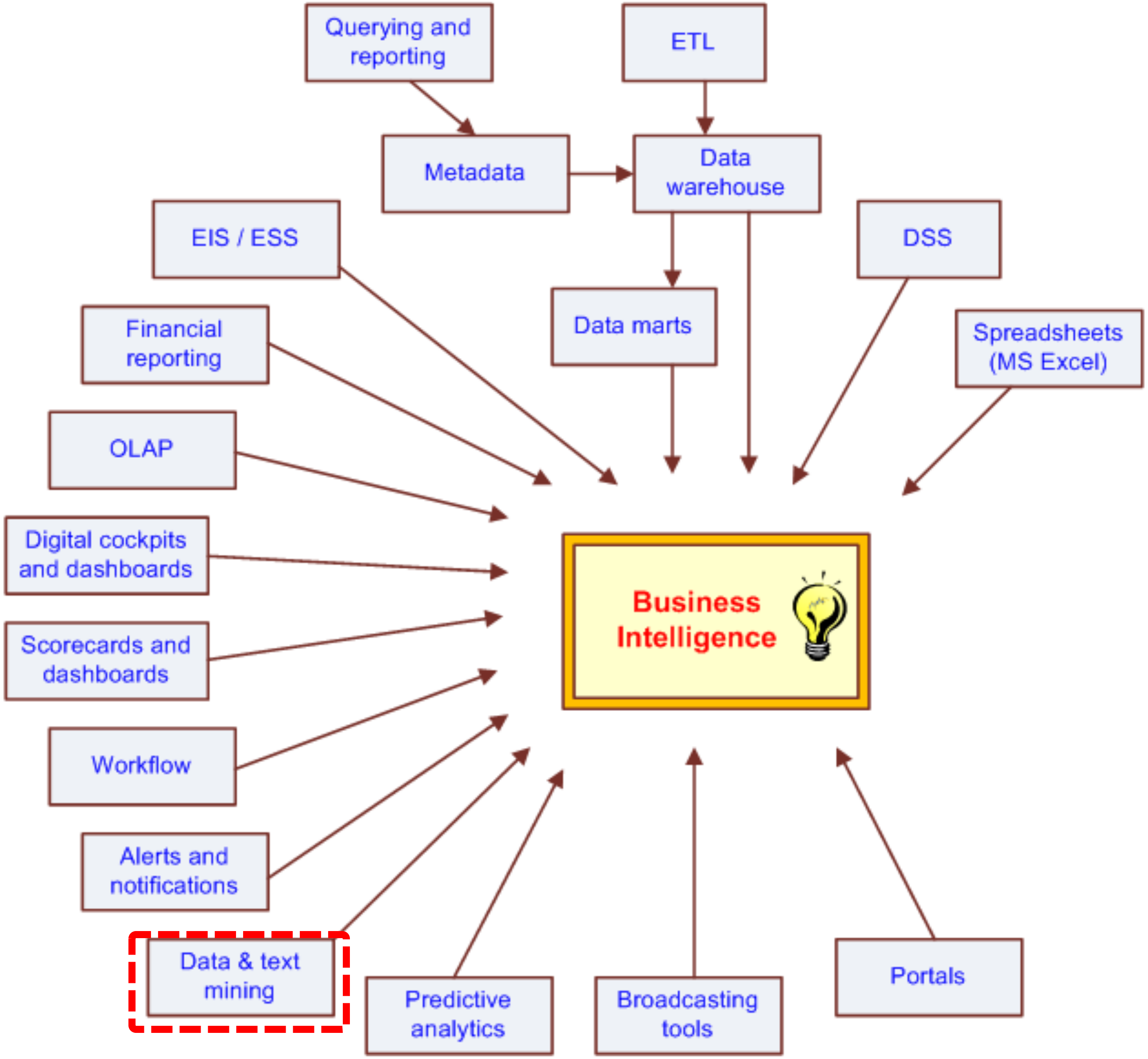
Business Analyst

Data Analyst

DBA

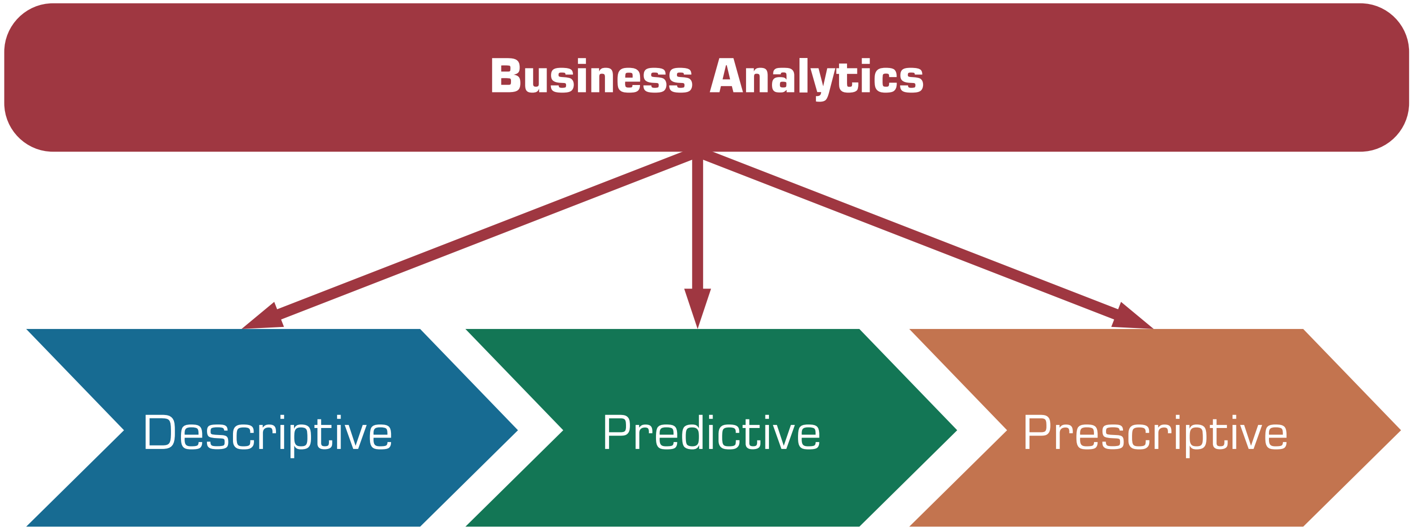
Source: Jiawei Han and Micheline Kamber (2006), Data Mining: Concepts and Techniques, Second Edition, Elsevier

The Evolution of BI Capabilities



Source: Turban et al. (2011), Decision Support and Business Intelligence Systems

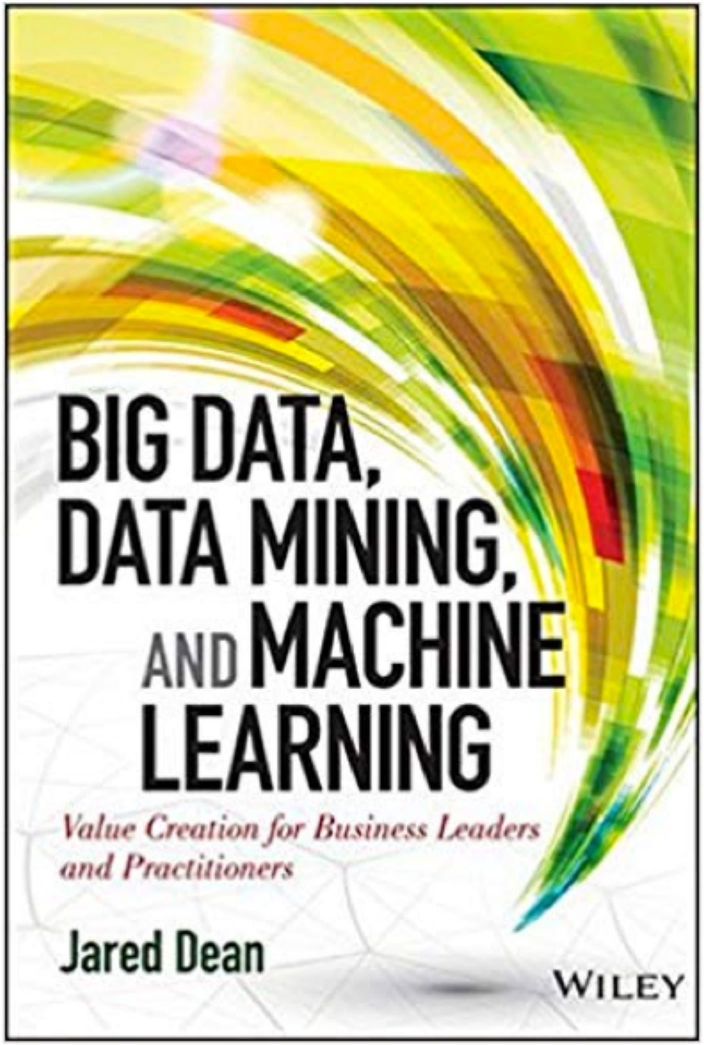
Three Types of Analytics



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

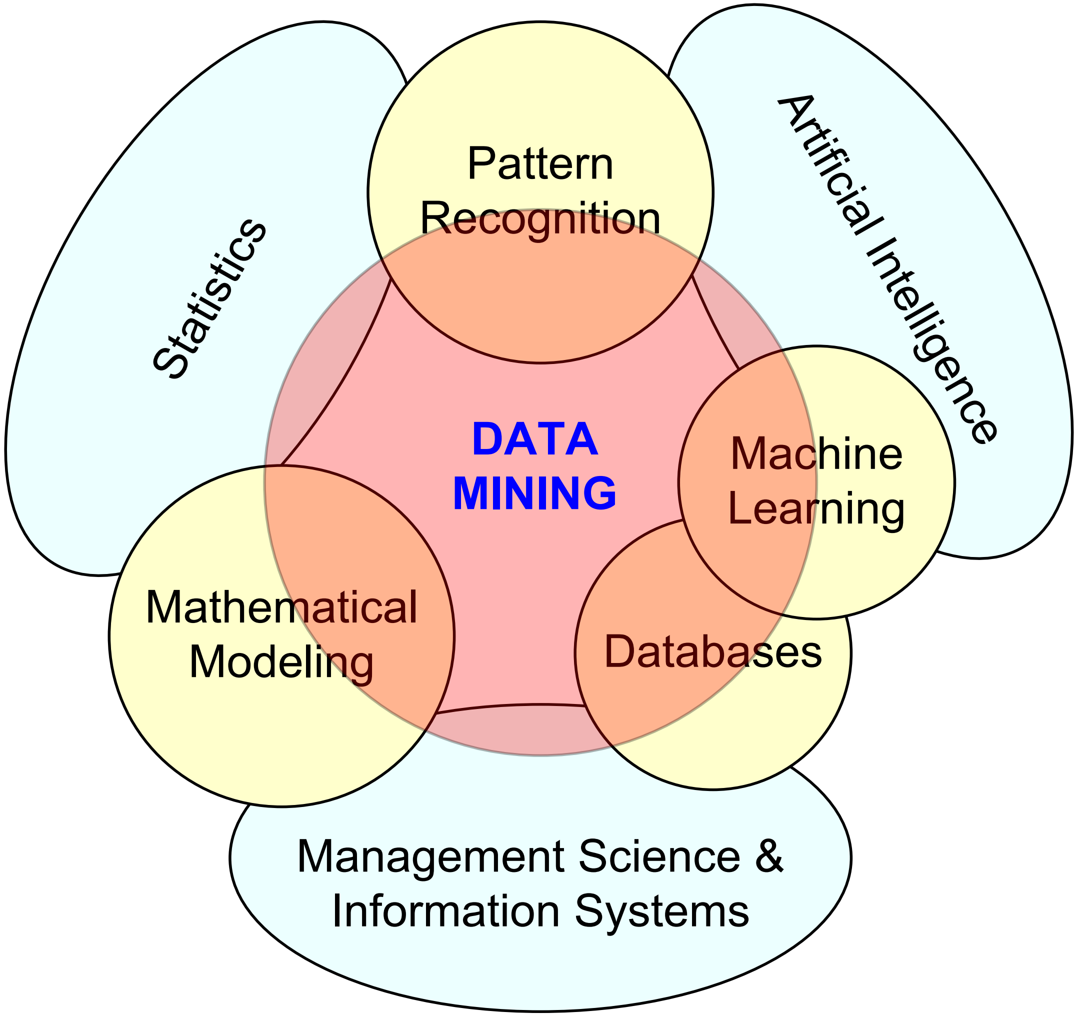
Source: Ramesh Sharda, Dursun Delen, and Efraim Turban (2017),
Business Intelligence, Analytics, and Data Science: A Managerial Perspective, 4th Edition, Pearson

Big Data, Data Mining, and Machine Learning: Value Creation for Business Leaders and Practitioners, Jared Dean, Wiley, 2014.

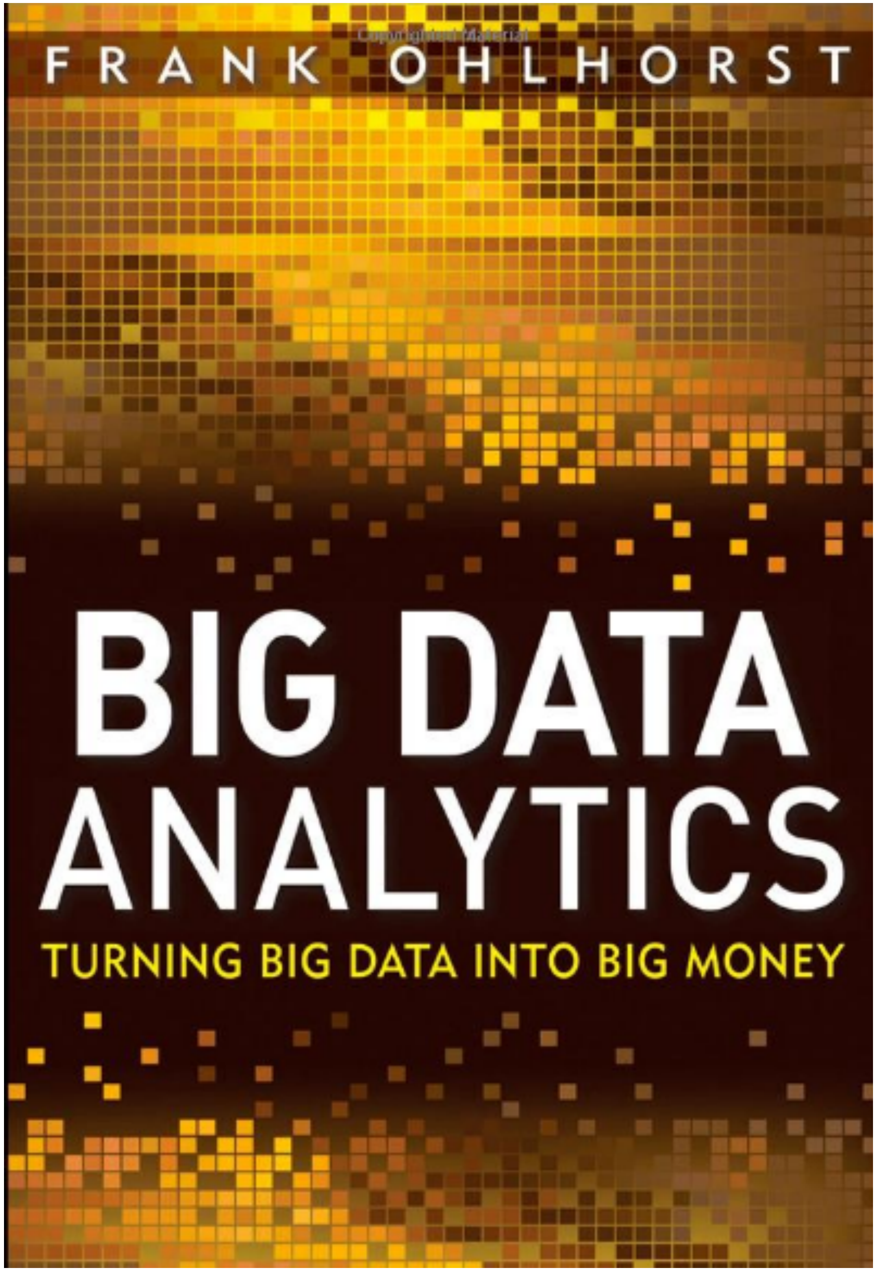


Source: https://www.amazon.com/Data-Mining-Machine-Learning-Practitioners/dp/1118618041

Data Mining at the Intersection of Many Disciplines



Source: Turban et al. (2011), Decision Support and Business Intelligence Systems



Copyrighted Material

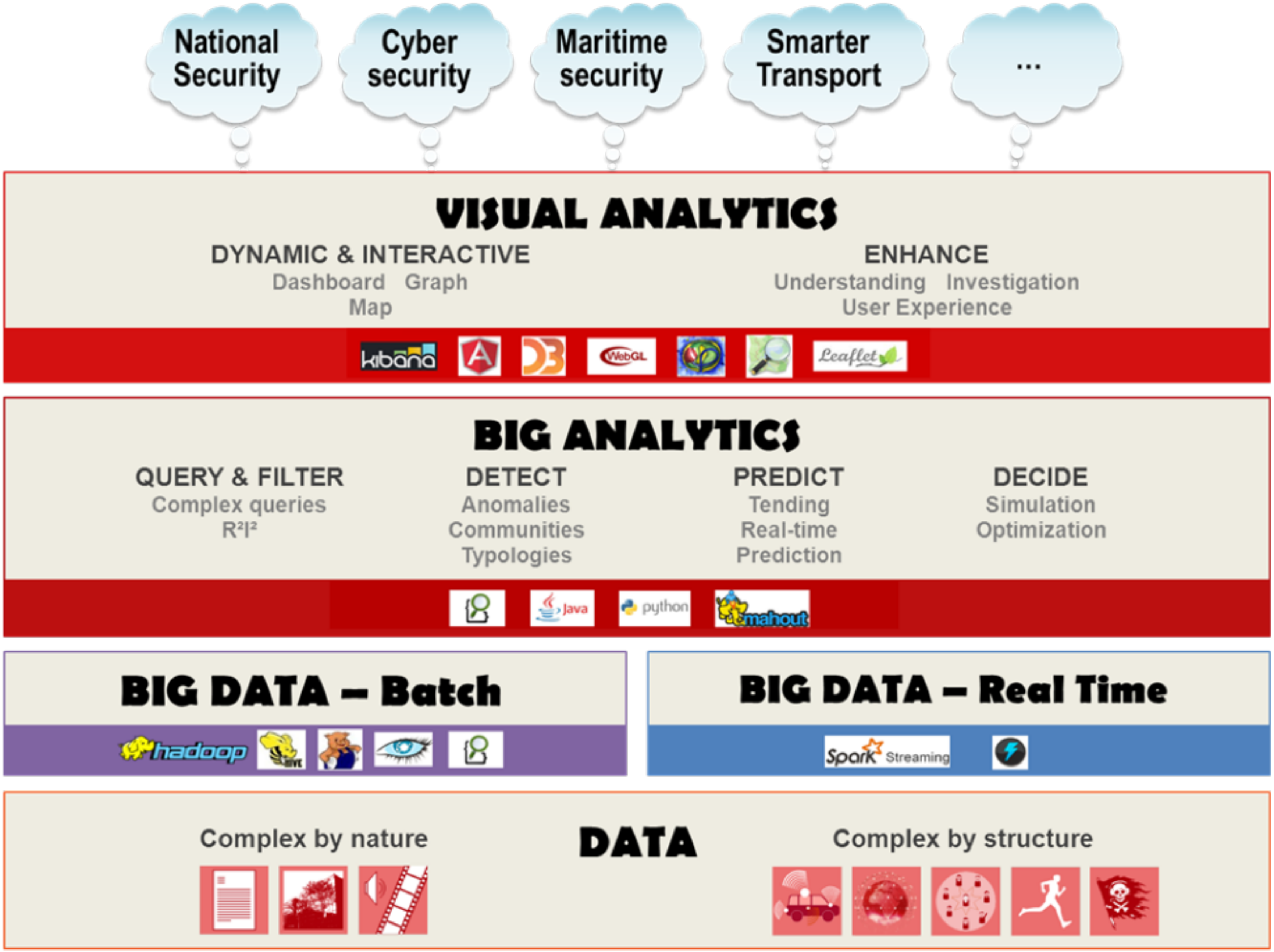
FRANK OHLHORST

BIG DATA ANALYTICS

TURNING BIG DATA INTO BIG MONEY

Source: http://www.amazon.com/Big-Data-Analytics-Turning-Money/dp/1118147596

Source: http://www.amazon.com/Big-Data-Revolution-Transform-Mayer-Schonberger/dp/B00D81X2YE



National Security

Cyber security

Maritime security

Smarter Transport

...

VISUAL ANALYTICS

DYNAMIC & INTERACTIVE

Dashboard Graph

Map

ENHANCE

Understanding Investigation

User Experience

BIG ANALYTICS

QUERY & FILTER

Complex queries

R²I²

DETECT

Anomalies

Communities

Typologies

PREDICT

Tending

Real-time

Prediction

DECIDE

Simulation

Optimization

BIG DATA – Batch

BIG DATA – Real Time

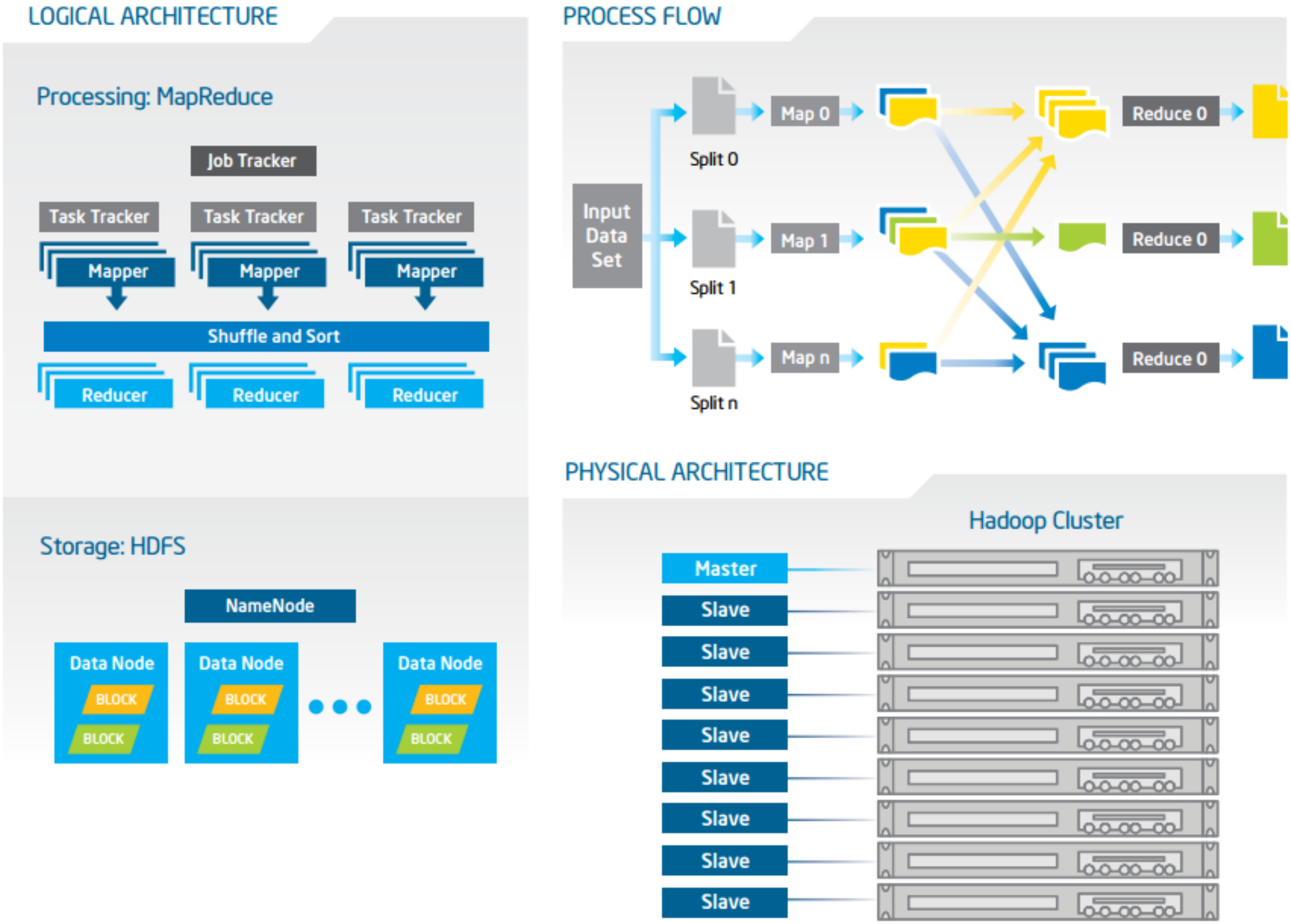
Complex by nature

DATA

Complex by structure

Source: https://www.thalesgroup.com/en/worldwide/big-data/big-data-big-analytics-visual-analytics-what-does-it-all-mean

Big Data with Hadoop Architecture

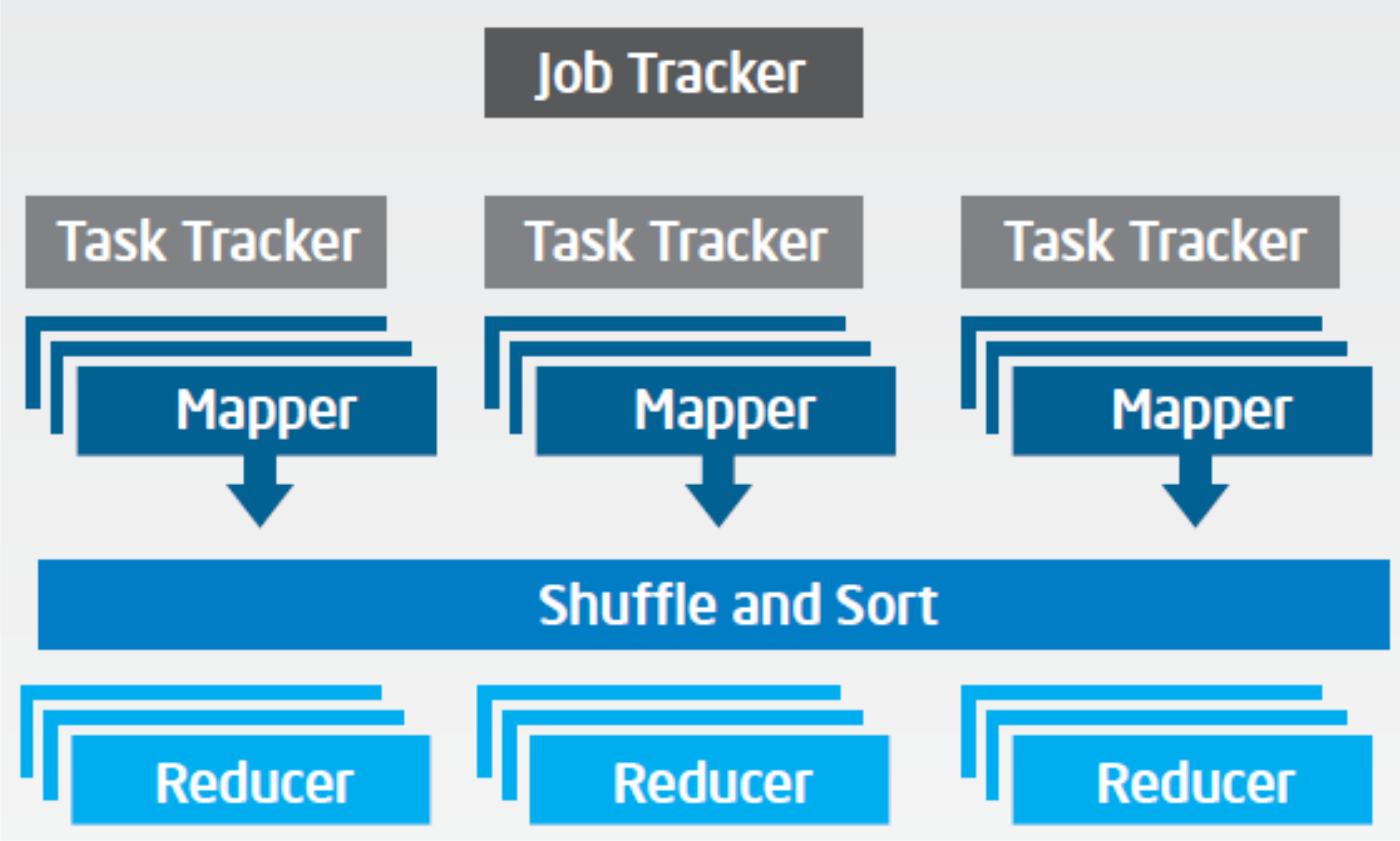


Source: https://software.intel.com/sites/default/files/article/402274/etl-big-data-with-hadoop.pdf

Big Data with Hadoop Architecture

Logical Architecture

Processing: MapReduce

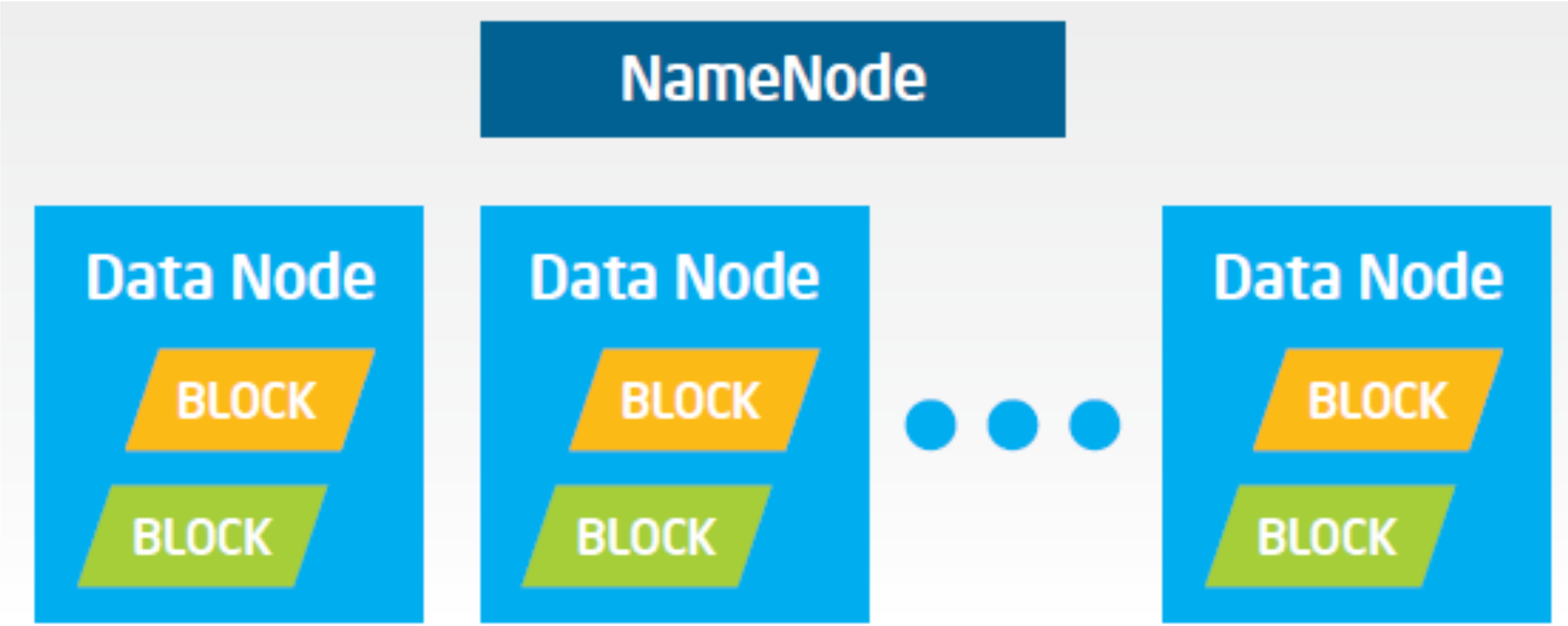


Source: https://software.intel.com/sites/default/files/article/402274/etl-big-data-with-hadoop.pdf

Big Data with Hadoop Architecture

Logical Architecture

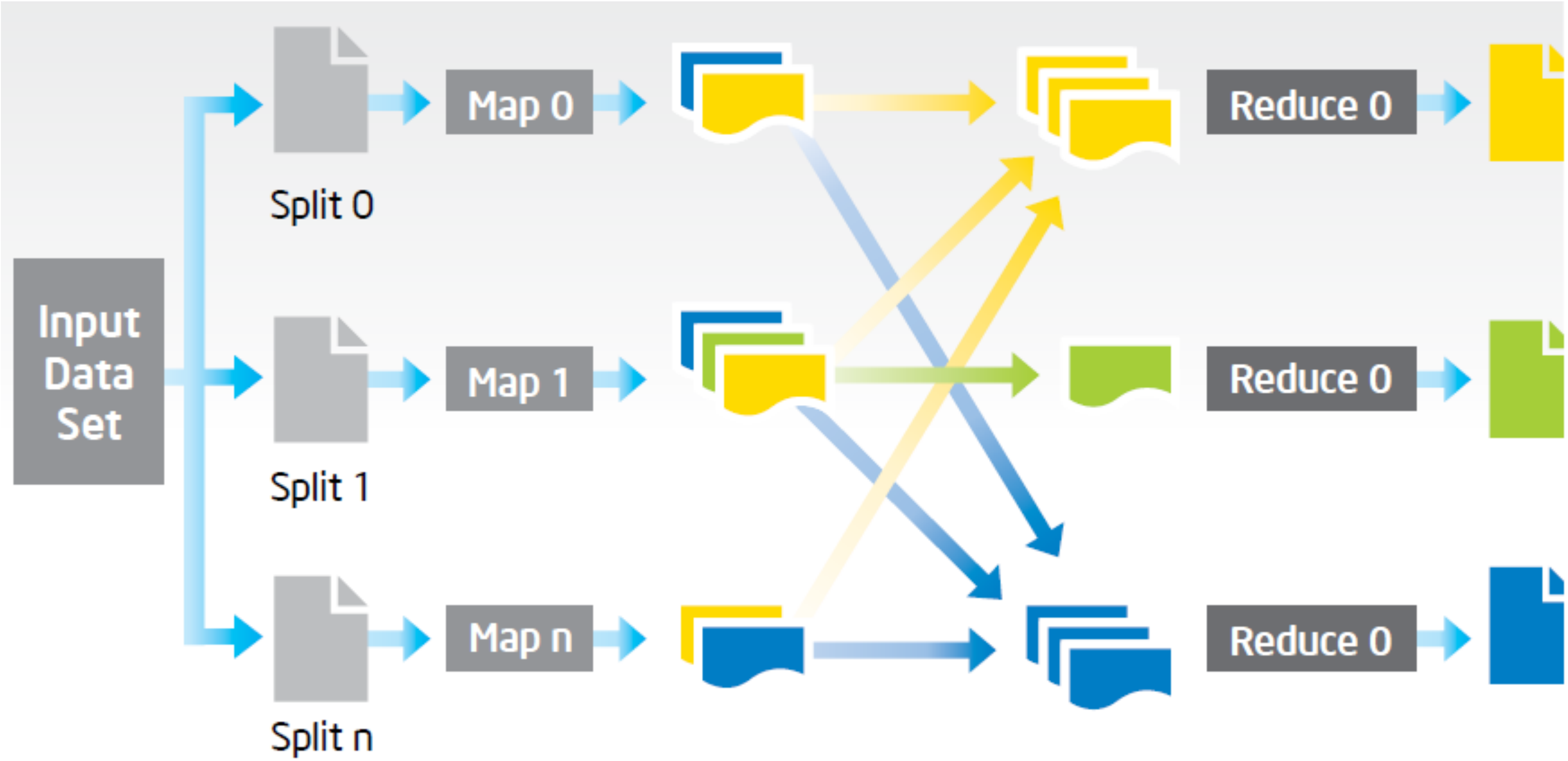
Storage: HDFS



Source: https://software.intel.com/sites/default/files/article/402274/etl-big-data-with-hadoop.pdf

Big Data with Hadoop Architecture

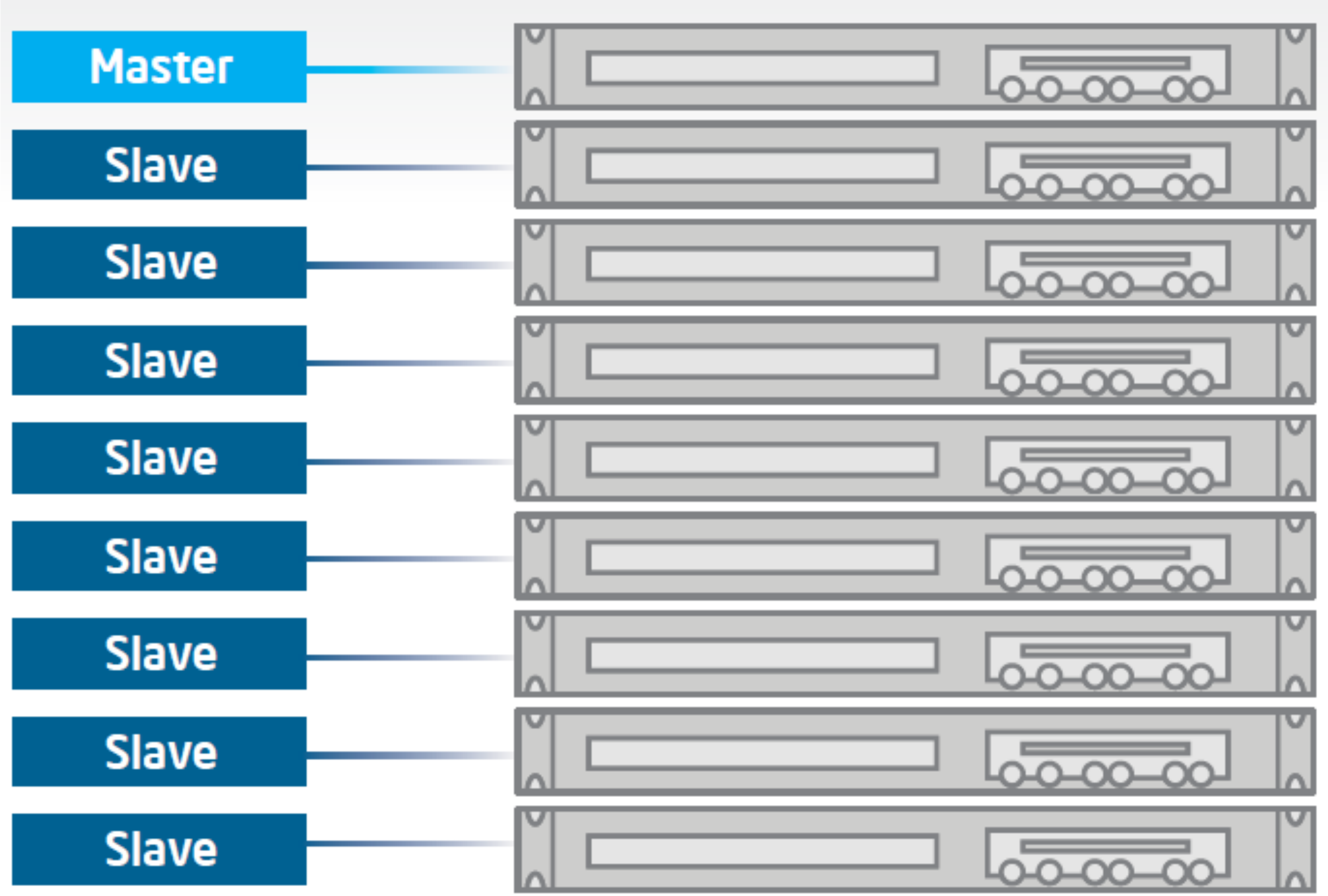
Process Flow



Source: https://software.intel.com/sites/default/files/article/402274/etl-big-data-with-hadoop.pdf

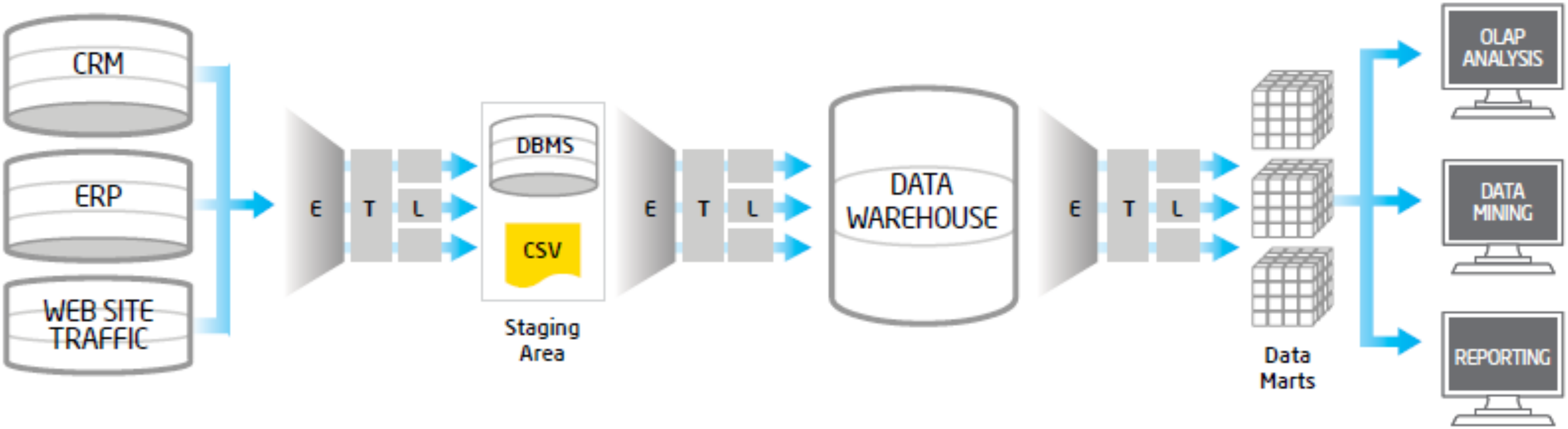
Big Data with Hadoop Architecture

Hadoop Cluster



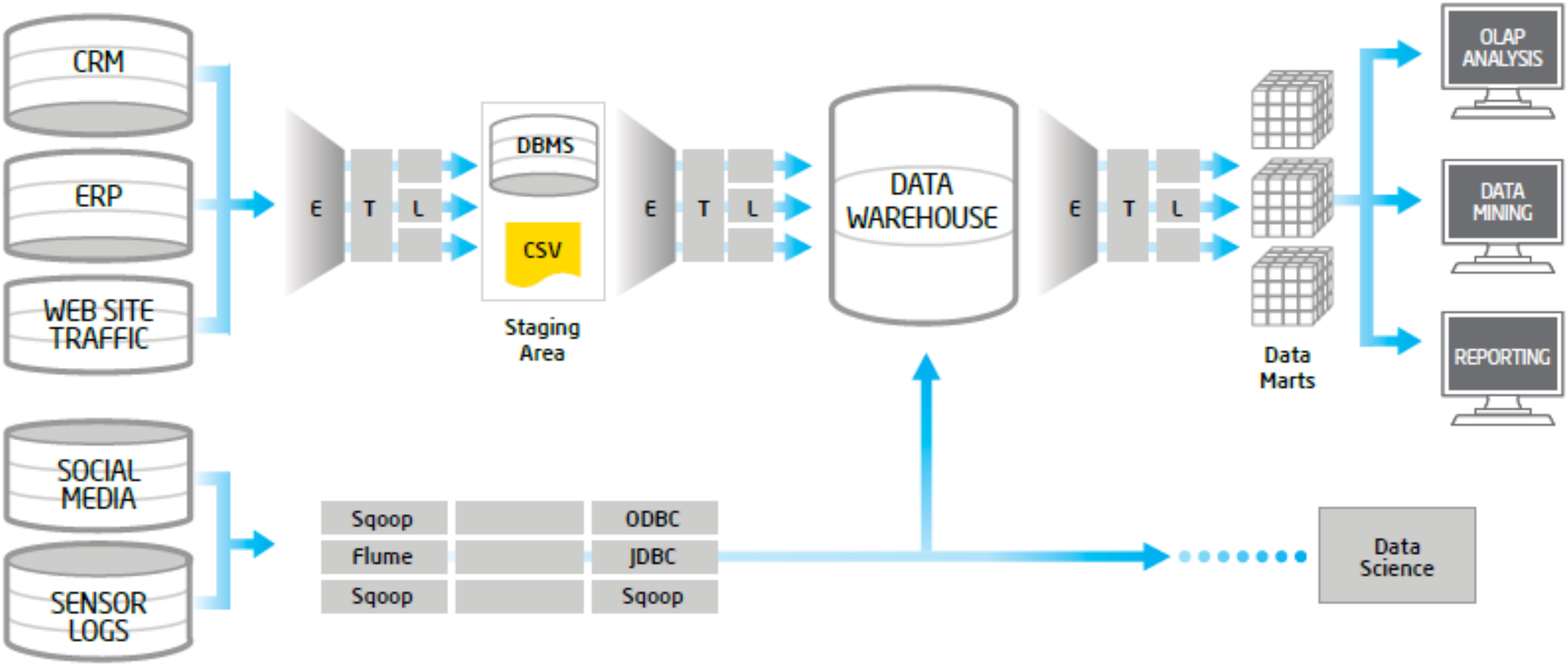
Source: https://software.intel.com/sites/default/files/article/402274/etl-big-data-with-hadoop.pdf

Traditional ETL Architecture



Source: https://software.intel.com/sites/default/files/article/402274/etl-big-data-with-hadoop.pdf

Offload ETL with Hadoop (Big Data Architecture)

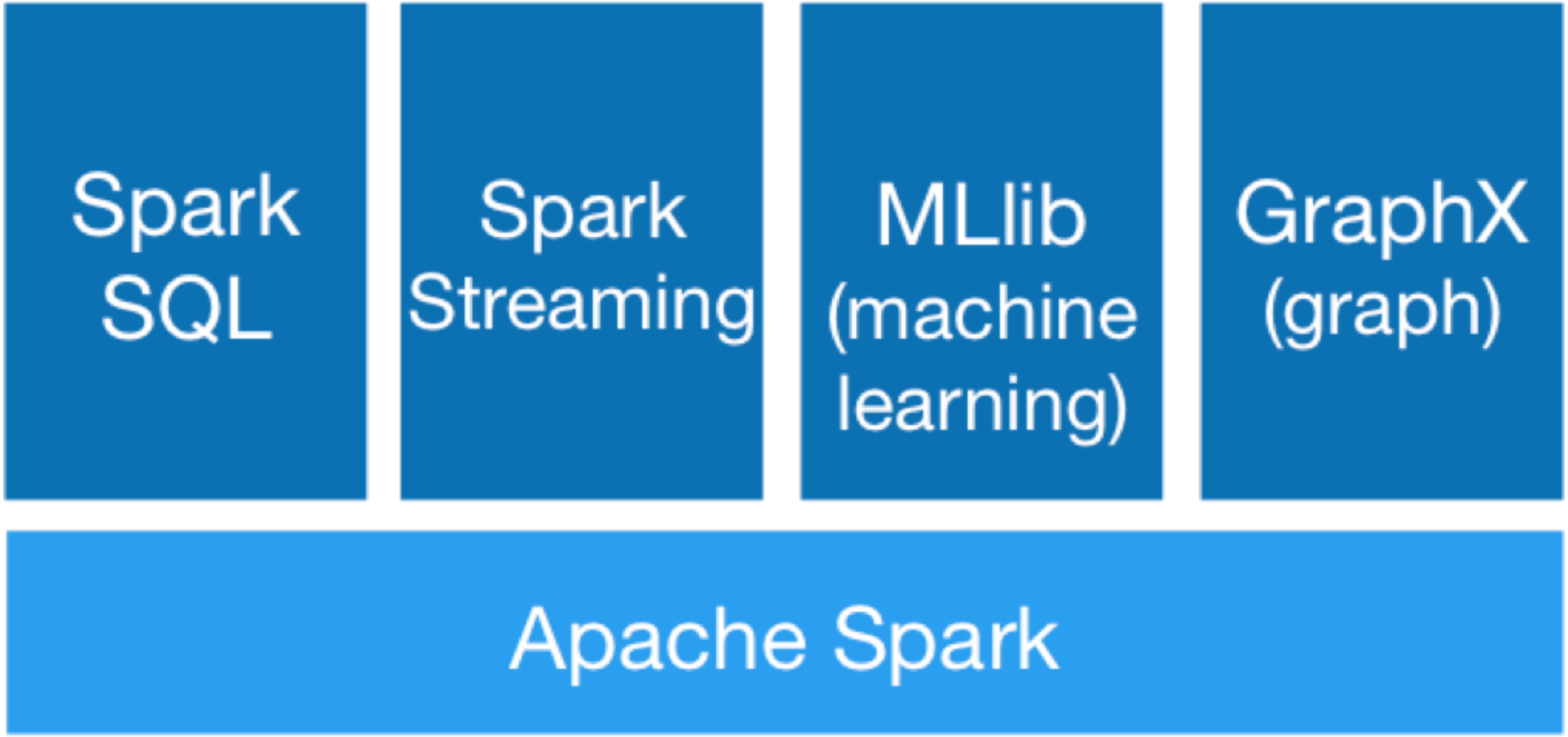


Source: https://software.intel.com/sites/default/files/article/402274/etl-big-data-with-hadoop.pdf

Spark and Hadoop

Source: http://spark.apache.org/

Spark Ecosystem



Spark SQL

Spark Streaming

MLlib (machine learning)

GraphX (graph)

Apache Spark

Source: http://spark.apache.org/

Summary

- This course introduces the fundamental concepts, research issues, and hands-on practices of big data analysis.
- Topics include
  1. Introduction to Big Data Analysis
  2. AI and Big Data Analysis
  3. Foundations of Big Data Analysis in Python
  4. Quantitative Big Data Analysis with Pandas in Python
  5. Machine Learning with Scikit-Learn In Python
  6. Deep Learning for Finance Big Data Analysis with TensorFlow
  7. Artificial Intelligence for Robo-Advisors
  8. Conversational Commerce and Intelligent Chatbots for Fintech
  9. Hands-on Practices with FintechSpace Digital Sandbox
  10. Case Study on Big Data Analysis

大數據分析
(Big Data Analysis)

國立臺北大學
NTPU
National Taipei University

Contact Information

aws academy
Accredited Educator

aws certified
Solutions Architect
Associate

aws certified
Cloud Practitioner

戴敏育 博士 (Min-Yuh Day, Ph.D.)

副教授 (Associate Professor)

國立臺北大學 資訊管理研究所

Institute of Information Management, National Taipei University

電話： 02-86741111 ext. 66873
研究室： 商8F12
地址： 23741 新北市三峽區大學路 151 號
Email：myday@gm.ntpu.edu.tw
網址：http://web.ntpu.edu.tw/~myday/