

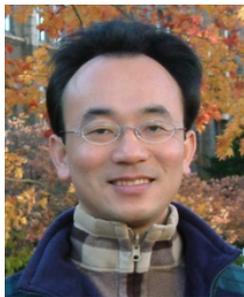
資料探勘 (Data Mining)

資料探勘介紹 (Introduction to data mining)

1092DM01

MBA, IM, NTPU (M5026) (Spring 2021)

Tue 2, 3, 4 (9:10-12:00) (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>

2021-02-23





戴敏育 博士 (Min-Yuh Day, Ph.D.)

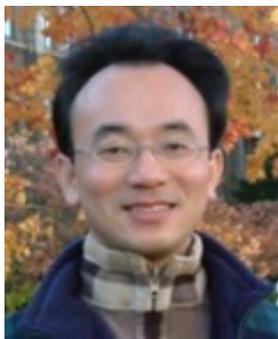
國立臺北大學 資訊管理研究所 副教授
中央研究院 資訊科學研究所 訪問學人
國立台灣大學 資訊管理 博士

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)





資料探勘

(Data Mining)

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/>



aws academy

Accredited
Educator



aws certified

Solutions
Architect

Associate



aws certified

Cloud
Practitioner

國立臺北大學

109學年度第2學期

課程大綱

Spring 2021 (2021.02 - 2021.06)

- 課程名稱：**資料探勘 (Data Mining)**
- 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：週二 2, 3, 4 (9:10-12:00)
- 上課教室：商8F40 (台北大學三峽校區)

教學目標

1. 瞭解資料探勘基本概念與研究議題。
2. 具備資料探勘實務操作能力。
3. 進行資料探勘相關之資訊管理研究。

Course Objectives

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

內容綱要

- 本課程介紹資料探勘基本概念、研究議題、與實務操作。
- 課程內容包括
 1. 資料探勘介紹
 2. ABC：人工智慧，大數據，雲端運算
 3. Python資料探勘的基礎
 4. 資料科學與資料探勘：發現，分析，可視化和呈現數據
 5. 非監督學習：關聯分析，購物籃分析
 6. 非監督學習：集群分析，行銷市場區隔
 7. 監督學習：分類和預測
 8. 機器學習和深度學習
 9. 卷積神經網絡、遞歸神經網絡、強化學習
 10. 社交網絡分析
 11. 資料探勘個案研究

Course Outline

- This course introduces the **fundamental concepts**, **research issues**, and **hands-on practices** of data mining.
- Topics include
 1. Introduction to data mining
 2. ABC: AI, Big Data, Cloud Computing
 3. Foundations of Data Mining in Python
 4. Data Science and Data Mining: Discovering, Analyzing, Visualizing and Presenting Data
 5. Unsupervised Learning: Association Analysis, Market Basket Analysis
 6. Unsupervised Learning: Cluster Analysis, Market Segmentation
 7. Supervised Learning: Classification and Prediction
 8. Machine Learning and Deep Learning
 9. Convolutional Neural Networks, Recurrent Neural Networks, Reinforcement Learning
 10. Social Network Analysis
 11. Case Study on Data Mining

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 | 2021/02/23 | 資料探勘介紹 (Introduction to data mining) |
| 2 | 2021/03/02 | ABC：人工智慧，大數據，雲端運算
(ABC: AI, Big Data, Cloud Computing) |
| 3 | 2021/03/09 | Python 資料探勘的基礎
(Foundations of Data Mining in Python) |
| 4 | 2021/03/16 | 資料科學與資料探勘：發現，分析，可視化和呈現數據
(Data Science and Data Mining:
Discovering, Analyzing, Visualizing and Presenting Data) |
| 5 | 2021/03/23 | 非監督學習：關聯分析，購物籃分析
(Unsupervised Learning: Association Analysis,
Market Basket Analysis) |
| 6 | 2021/03/30 | 資料探勘個案研究 I
(Case Study on Data Mining I) |

課程大綱 (Syllabus)

- | 週次 (Week) | 日期 (Date) | 內容 (Subject/Topics) |
|-----------|------------|---|
| 7 | 2021/04/06 | 非監督學習：集群分析，行銷市場區隔
(Unsupervised Learning: Cluster Analysis, Market Segmentation) |
| 8 | 2021/04/13 | 監督學習：分類和預測
(Supervised Learning: Classification and Prediction) |
| 9 | 2021/04/20 | 期中報告 (Midterm Project Report) |
| 10 | 2021/04/27 | 監督學習：分類和預測
(Supervised Learning: Classification and Prediction) |
| 11 | 2021/05/04 | 機器學習和深度學習
(Machine Learning and Deep Learning) |
| 12 | 2021/05/11 | 卷積神經網絡
(Convolutional Neural Networks) |

課程大綱 (Syllabus)

週次 (Week)	日期 (Date)	內容 (Subject/Topics)
13	2021/05/18	資料探勘個案研究 II (Case Study on Data Mining II)
14	2021/05/25	遞歸神經網絡 (Recurrent Neural Networks)
15	2021/06/01	強化學習 (Reinforcement Learning)
16	2021/06/08	社交網絡分析 (Social Network Analysis)
17	2021/06/15	期末報告 I (Final Project Report I)
18	2021/06/22	期末報告 II (Final Project Report II)

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)

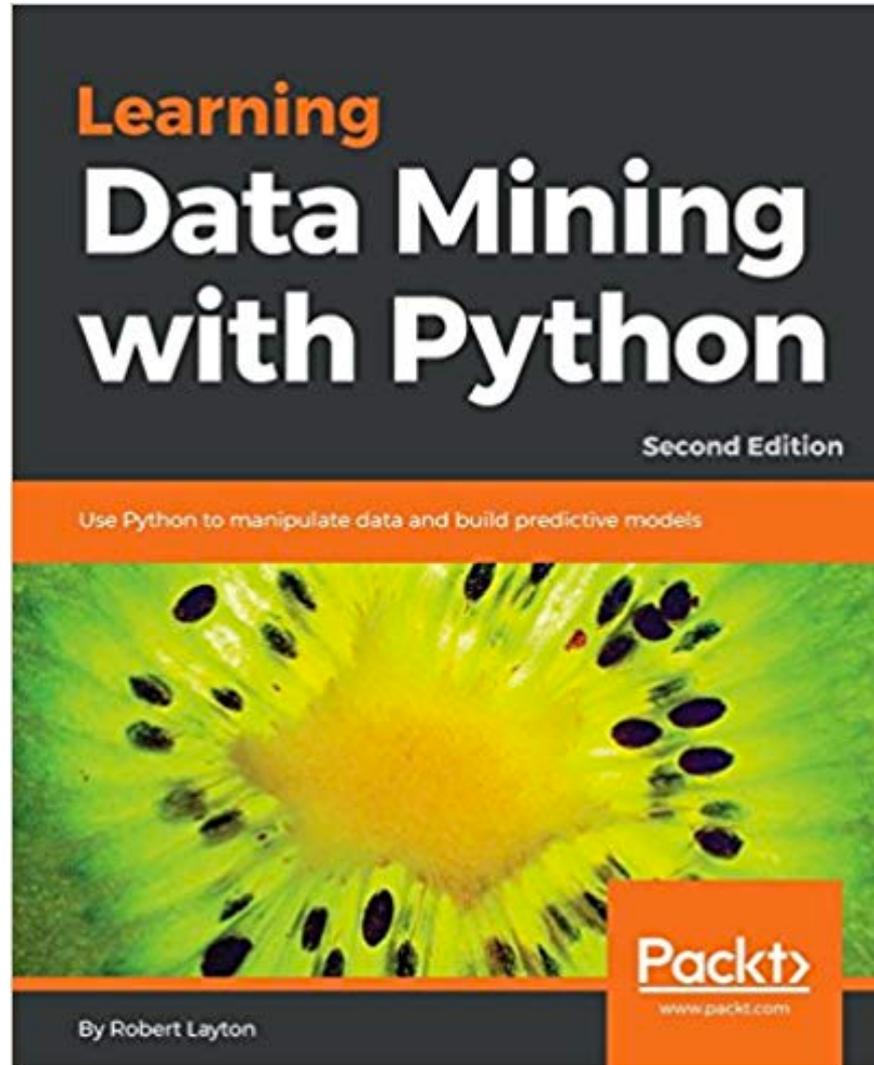
- Robert Layton (2017),
Learning Data Mining with Python,
Second Edition, Packt Publishing.

參考書目

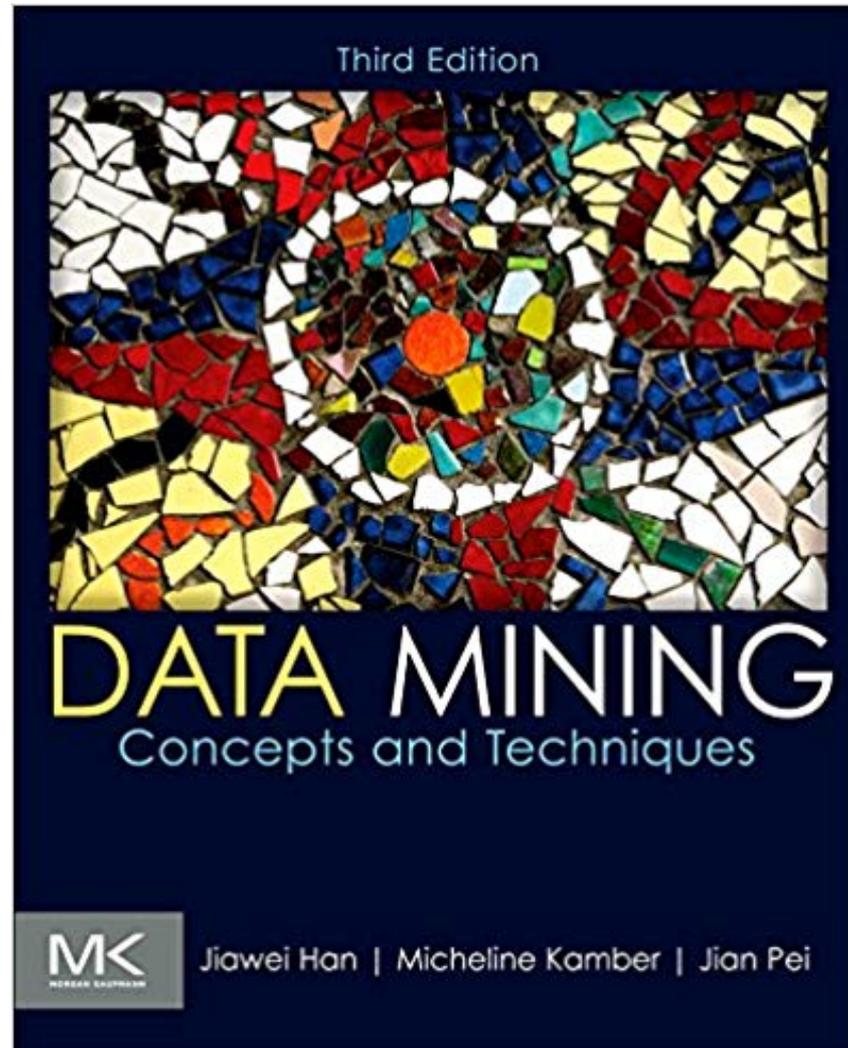
(Reference Books)

- 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.

Learning Data Mining with Python - Second Edition,
Robert Layton,
Packt Publishing, 2017

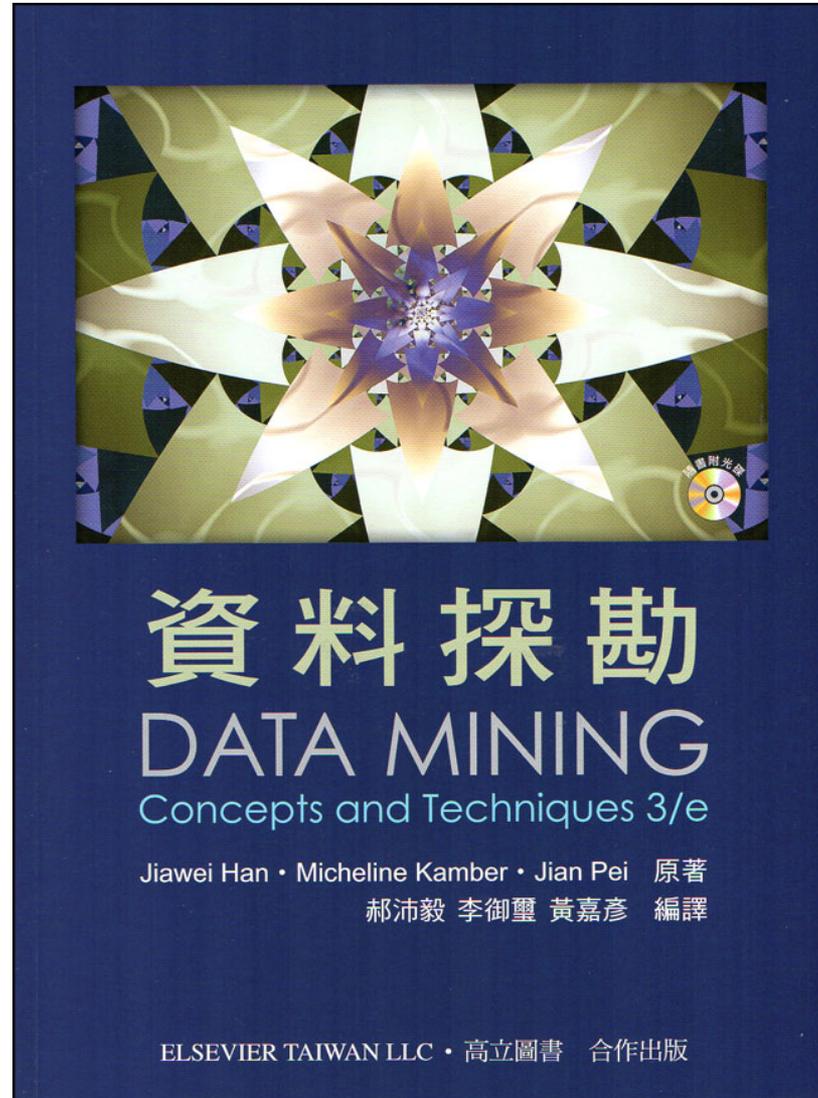


**Data Mining: Concepts and Techniques, Third Edition,
Jiawei Han, Micheline Kamber and Jian Pei,
Morgan Kaufmann, 2011**

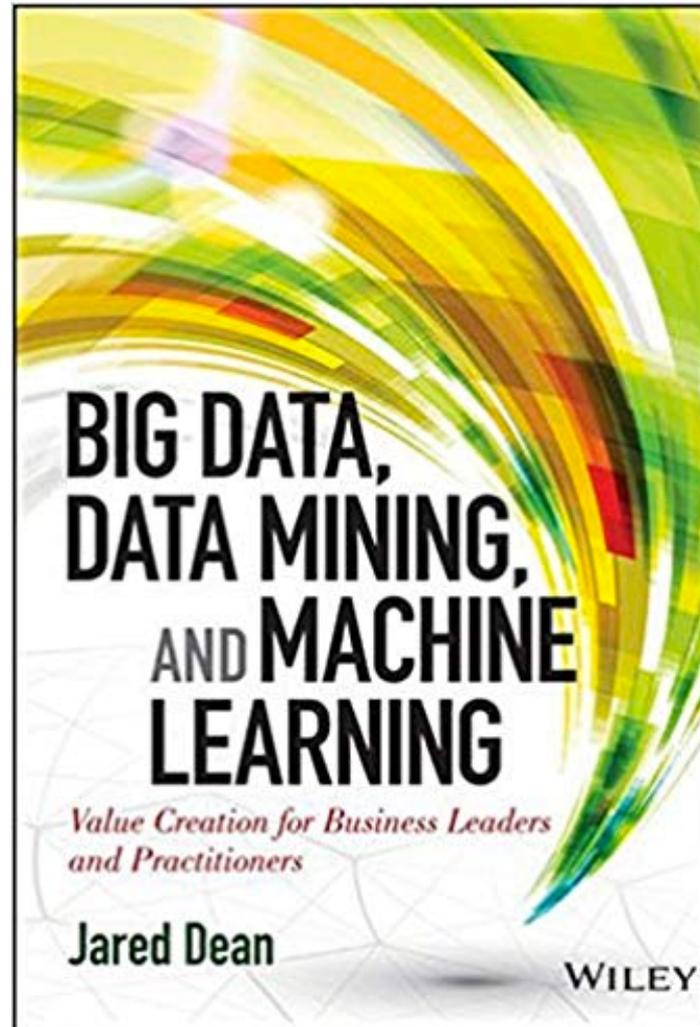


郝沛毅, 李御璽, 黃嘉彥 編譯, 資料探勘

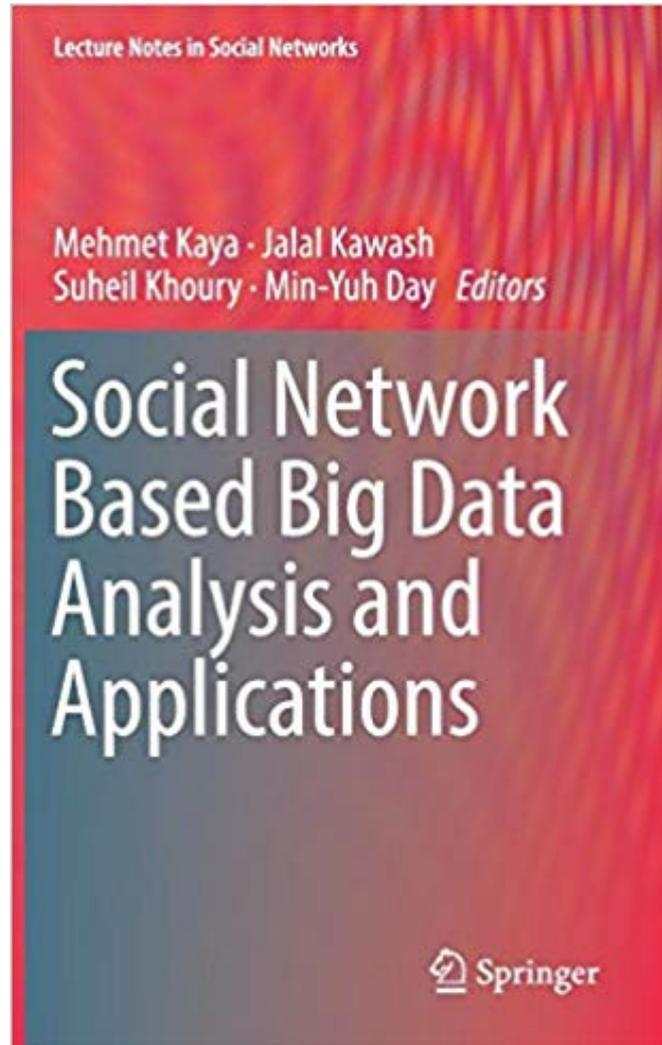
(Jiawei Han, Micheline Kamber, Jian Pei, Data Mining - Concepts and Techniques 3/e),
高立圖書, 2014



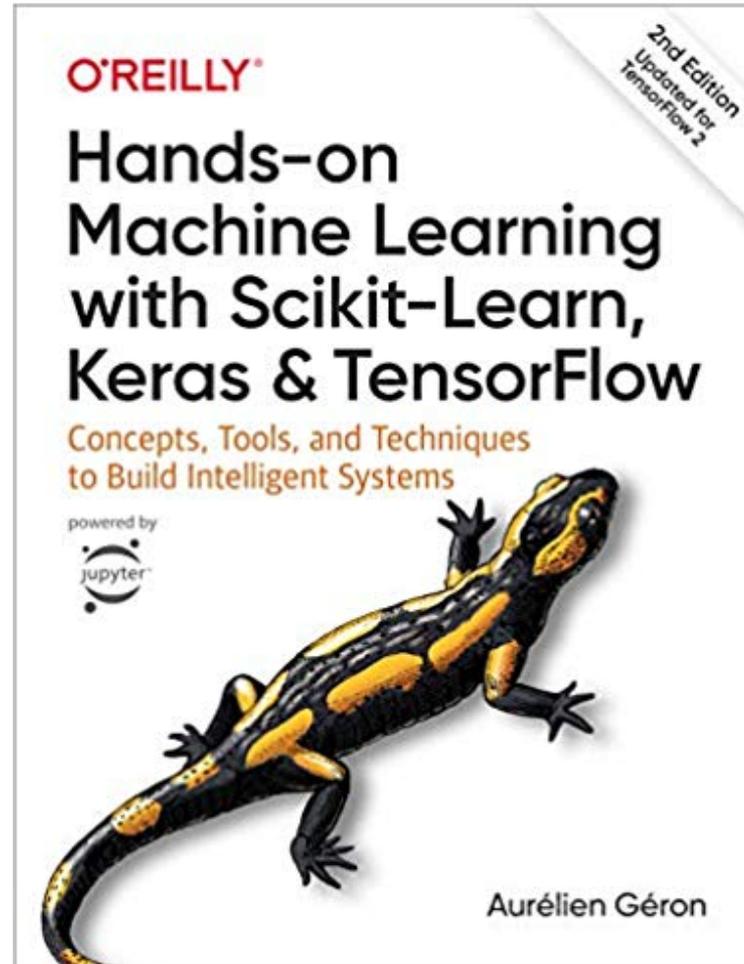
**Big Data, Data Mining, and Machine Learning: Value Creation for
Business Leaders and Practitioners,
Jared Dean,
Wiley, 2014.**



**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.**



**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**

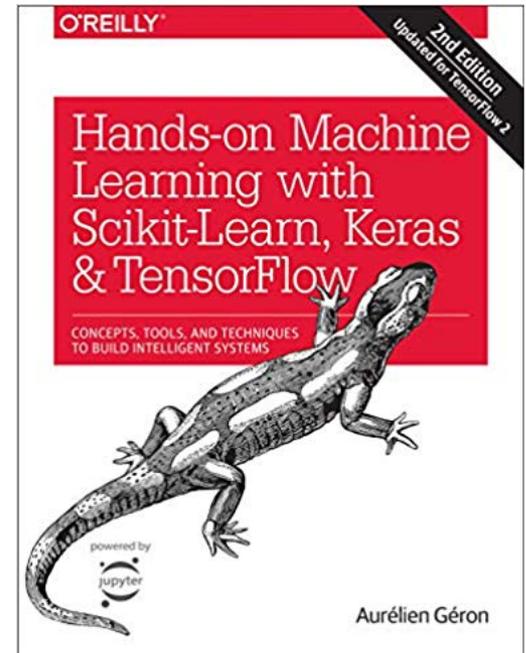


<https://github.com/ageron/handson-ml2>

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

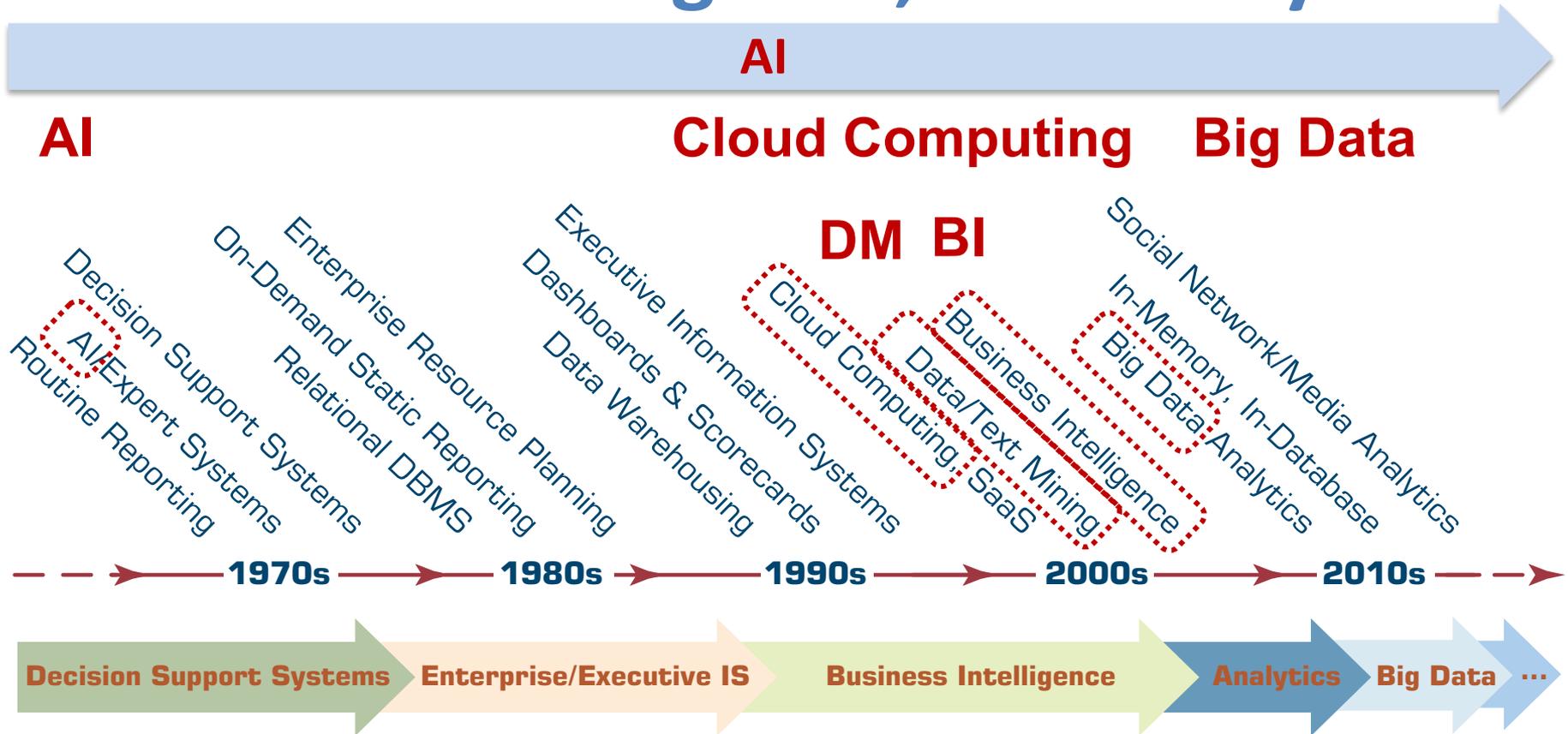
Notebooks

- [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](#)



AI, Big Data, Cloud Computing

Evolution of Decision Support, Business Intelligence, and Analytics

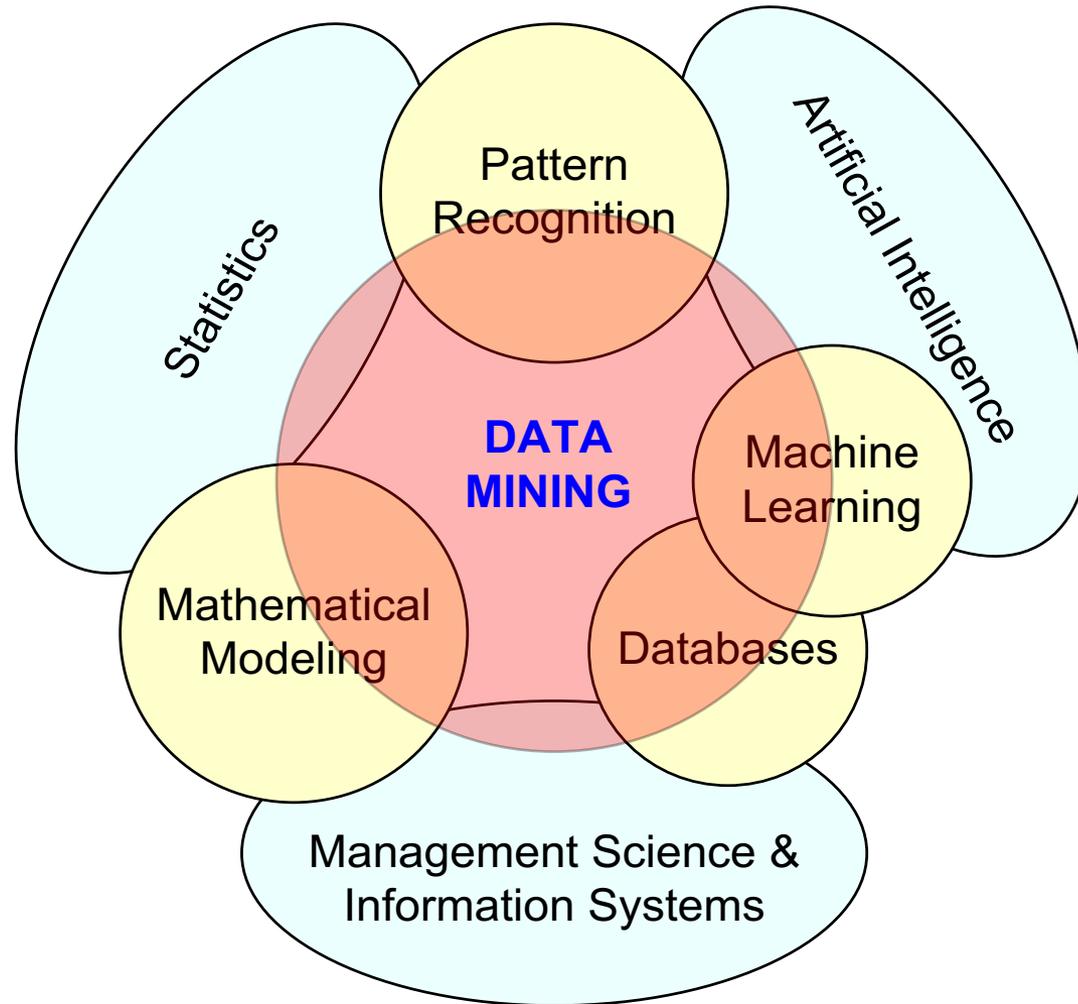


Data Mining

Is a Blend of Multiple Disciplines



Data Mining at the Intersection of Many Disciplines



Data Mining Tasks & Methods

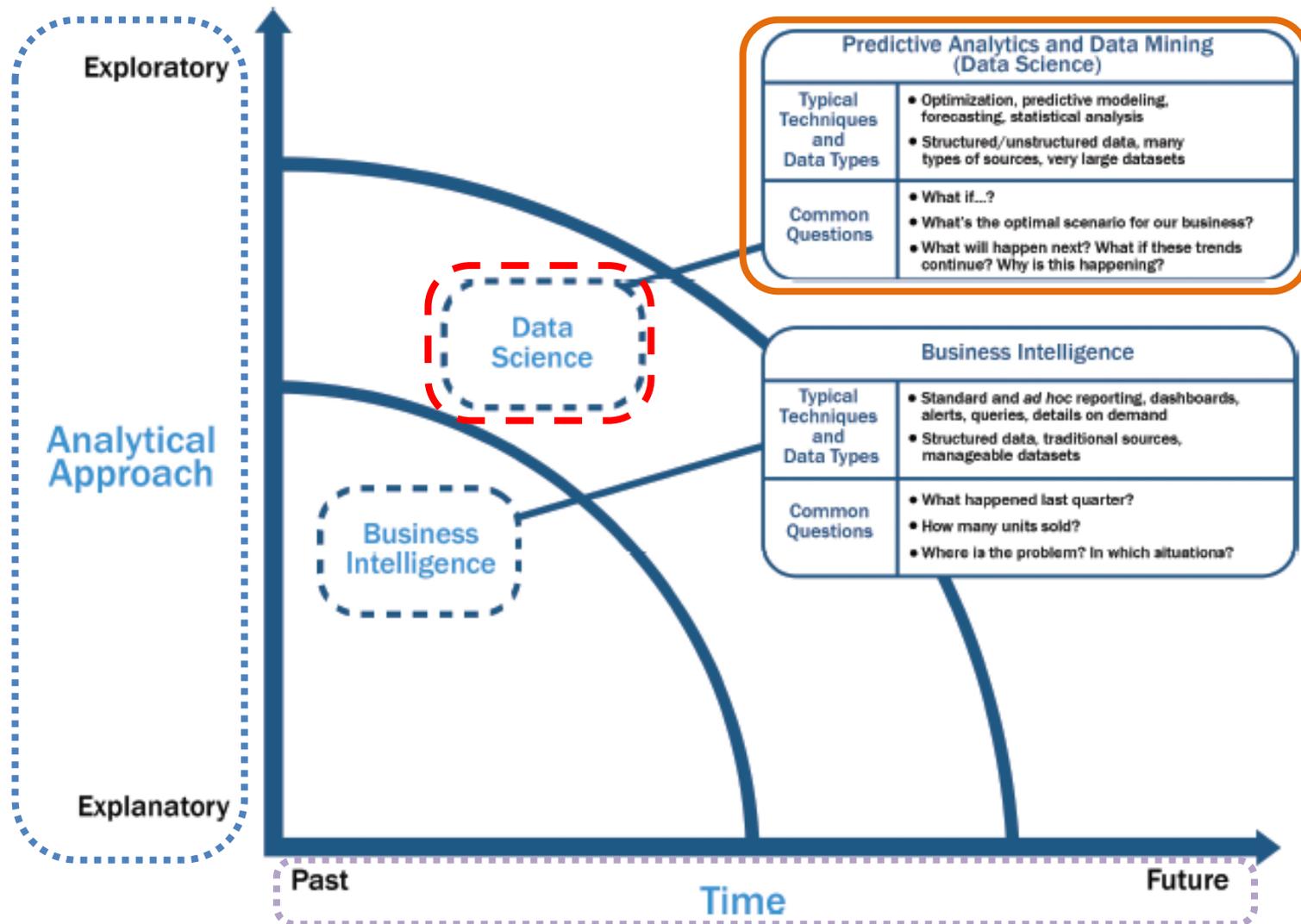
Prediction

Association

Segmentation

Data Mining Tasks & Methods	Data Mining Algorithms	Learning Type
Prediction		
Classification	Decision Trees, Neural Networks, Support Vector Machines, kNN, Naïve Bayes, GA	Supervised
Regression	Linear/Nonlinear Regression, ANN, Regression Trees, SVM, kNN, GA	Supervised
Time series	Autoregressive Methods, Averaging Methods, Exponential Smoothing, ARIMA	Supervised
Association		
Market-basket	Apriori, OneR, ZeroR, Eclat, GA	Unsupervised
Link analysis	Expectation Maximization, Apriori Algorithm, Graph-Based Matching	Unsupervised
Sequence analysis	Apriori Algorithm, FP-Growth, Graph-Based Matching	Unsupervised
Segmentation		
Clustering	k-means, Expectation Maximization (EM)	Unsupervised
Outlier analysis	k-means, Expectation Maximization (EM)	Unsupervised

Data Science and Business Intelligence



Data Science and Business Intelligence



Predictive Analytics and Data Mining (Data Science)

Past

Time

Future

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?



Data Mining:

Core **Analytics** Process

The **KDD** Process for
Extracting Useful **Knowledge**
from Volumes of **Data**

The **KDD Process** for Extracting Useful **Knowledge** from Volumes of **Data**.

Communications of the ACM, 39(11), 27-34.

Knowledge Discovery in Databases creates the context for developing the tools needed to control the flood of data facing organizations that depend on ever-growing databases of business, manufacturing, scientific, and personal information.

The KDD Process for Extracting Useful Knowledge from Volumes of Data

AS WE MARCH INTO THE AGE of digital information, the problem of data overload looms ominously ahead. Our ability to analyze and understand massive datasets lags far behind our ability to gather and store the data. A new generation of computational techniques and tools is required to support the extraction of useful knowledge from the rapidly growing volumes of data. These techniques and tools are the subject of the emerging field of knowledge discovery in databases (KDD) and data mining.

Large databases of digital information are ubiquitous. Data from the neighborhood store's checkout register, your bank's credit card authorization device, records in your doctor's office, patterns in your telephone calls,

and many more applications generate streams of digital records archived in huge databases, sometimes in so-called data warehouses.

Current hardware and database technology allow efficient and inexpensive reliable data storage and access. However, whether the context is business, medicine, science, or government, the datasets themselves (in raw form) are of little direct value. What is of value is the knowledge that can be inferred from the data and put to use. For example, the marketing database of a consumer

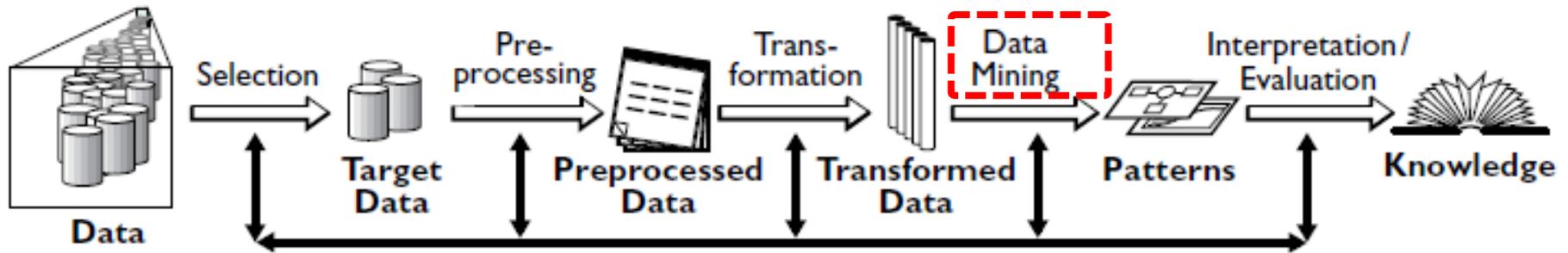
Usama Fayyad,
Gregory Piatetsky-Shapiro,
and Padhraic Smyth



Data Mining

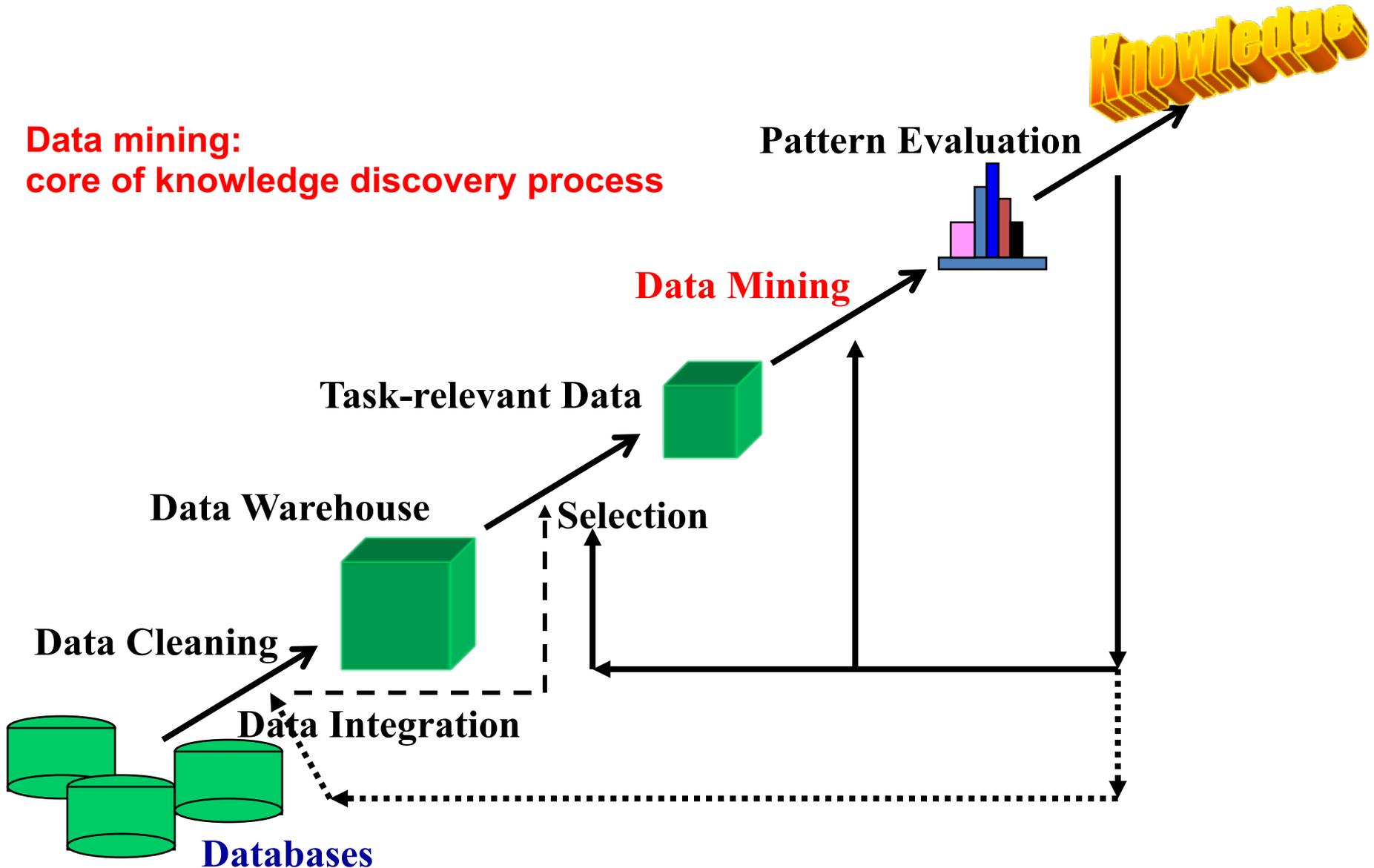
Knowledge Discovery in Databases (KDD) Process

(Fayyad et al., 1996)



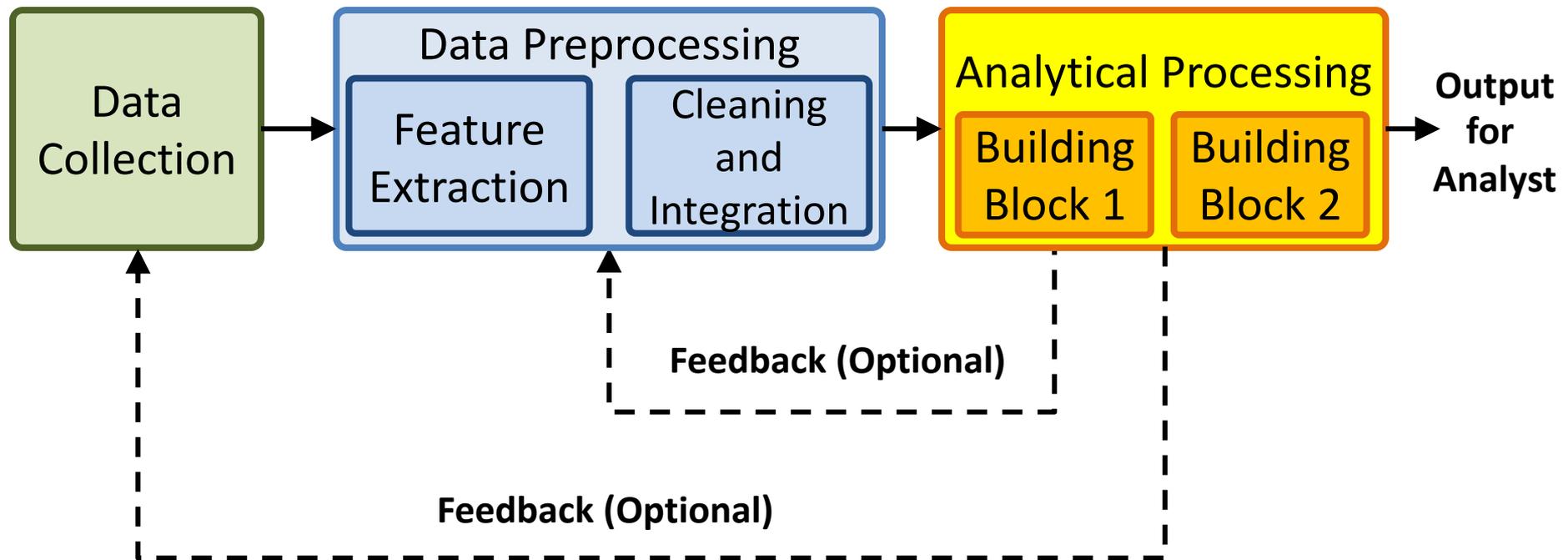
Knowledge Discovery (KDD) Process

Data mining:
core of knowledge discovery process

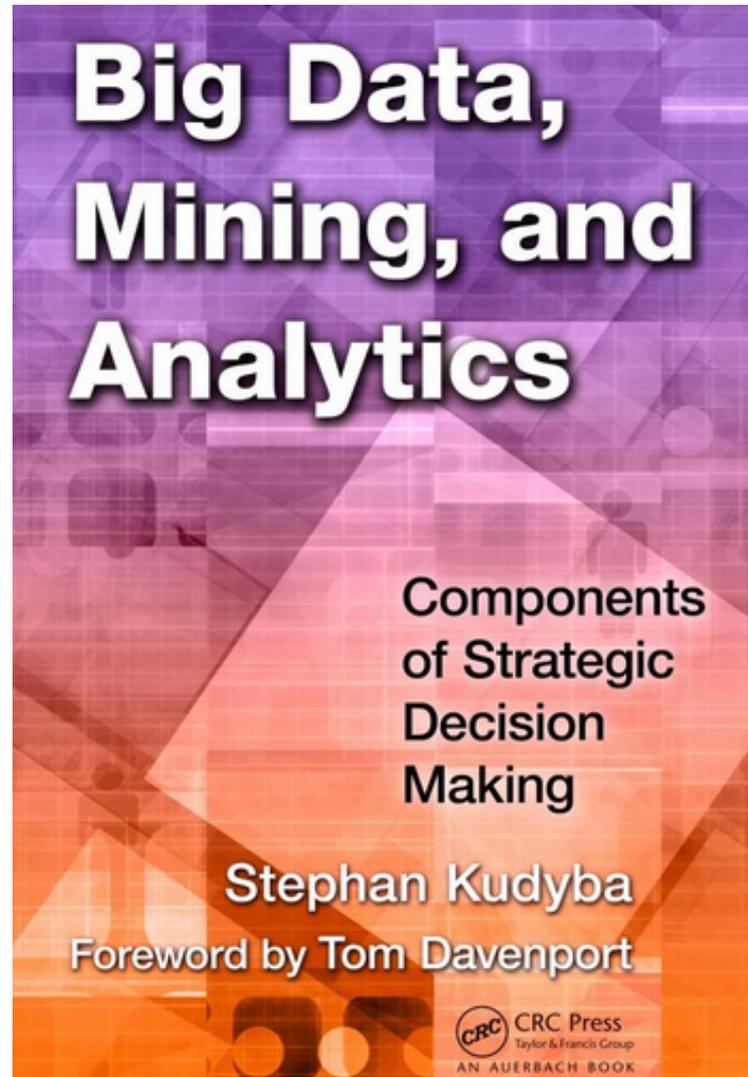


Data Mining Processing Pipeline

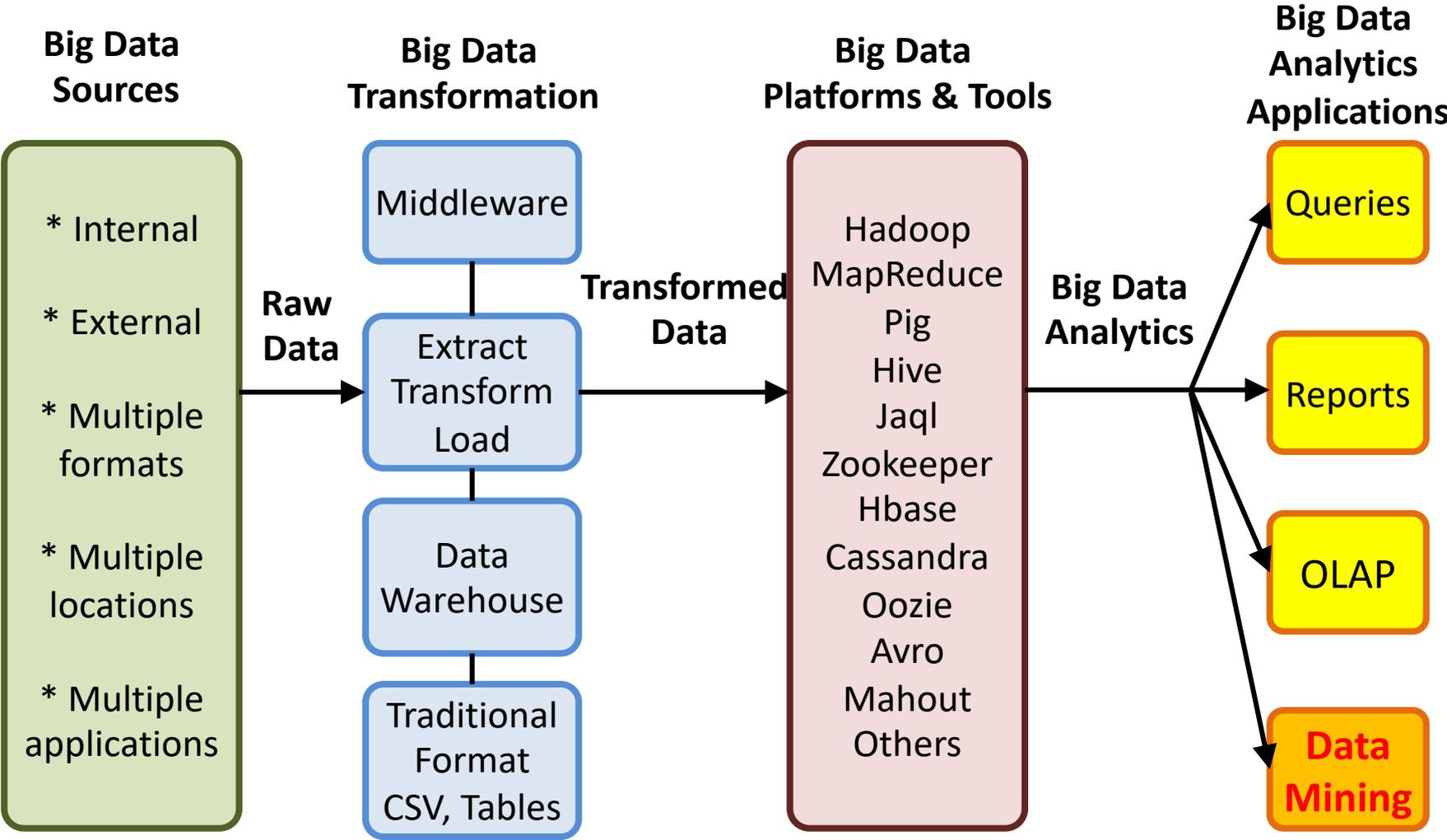
(Charu Aggarwal, 2015)



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



Architecture of Big Data Analytics



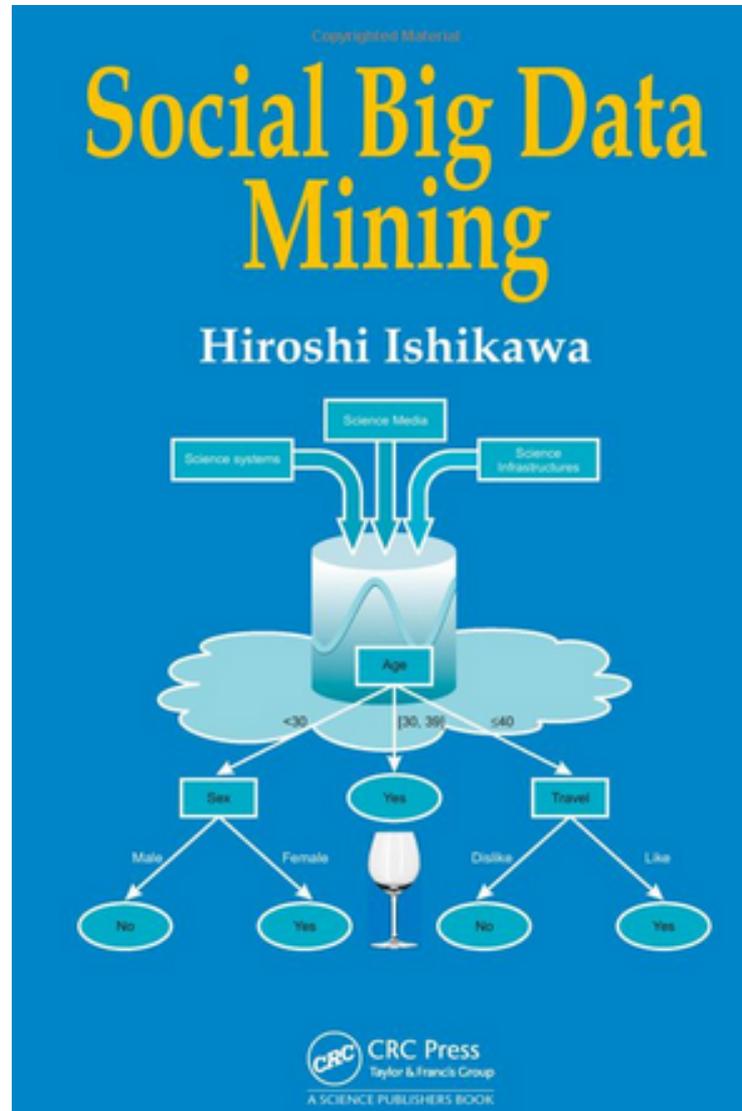
Source: Stephan Kudyba (2014), Big Data, Mining, and Analytics: Components of Strategic Decision Making, Auerbach Publications

Architecture of Big Data Analytics



Social Big Data Mining

(Hiroshi Ishikawa, 2015)



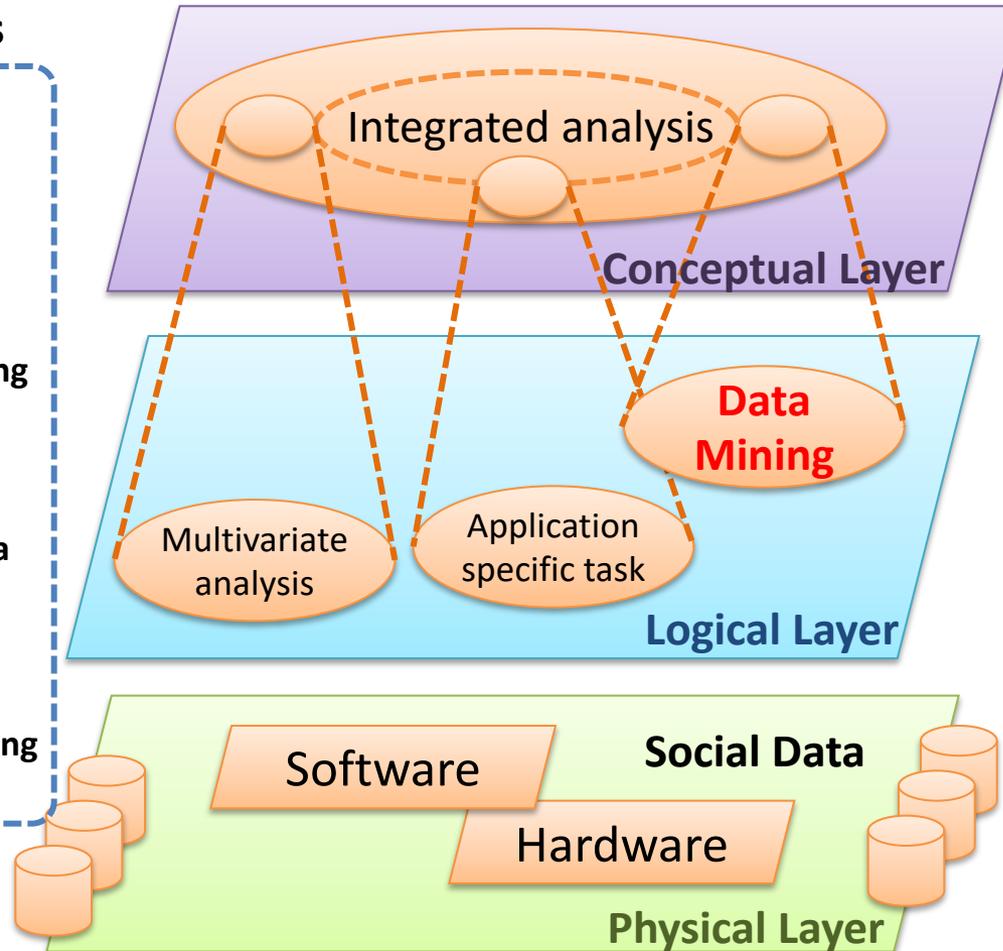
Source: <http://www.amazon.com/Social-Data-Mining-Hiroshi-Ishikawa/dp/149871093X>

Architecture for Social Big Data Mining

(Hiroshi Ishikawa, 2015)

Enabling Technologies

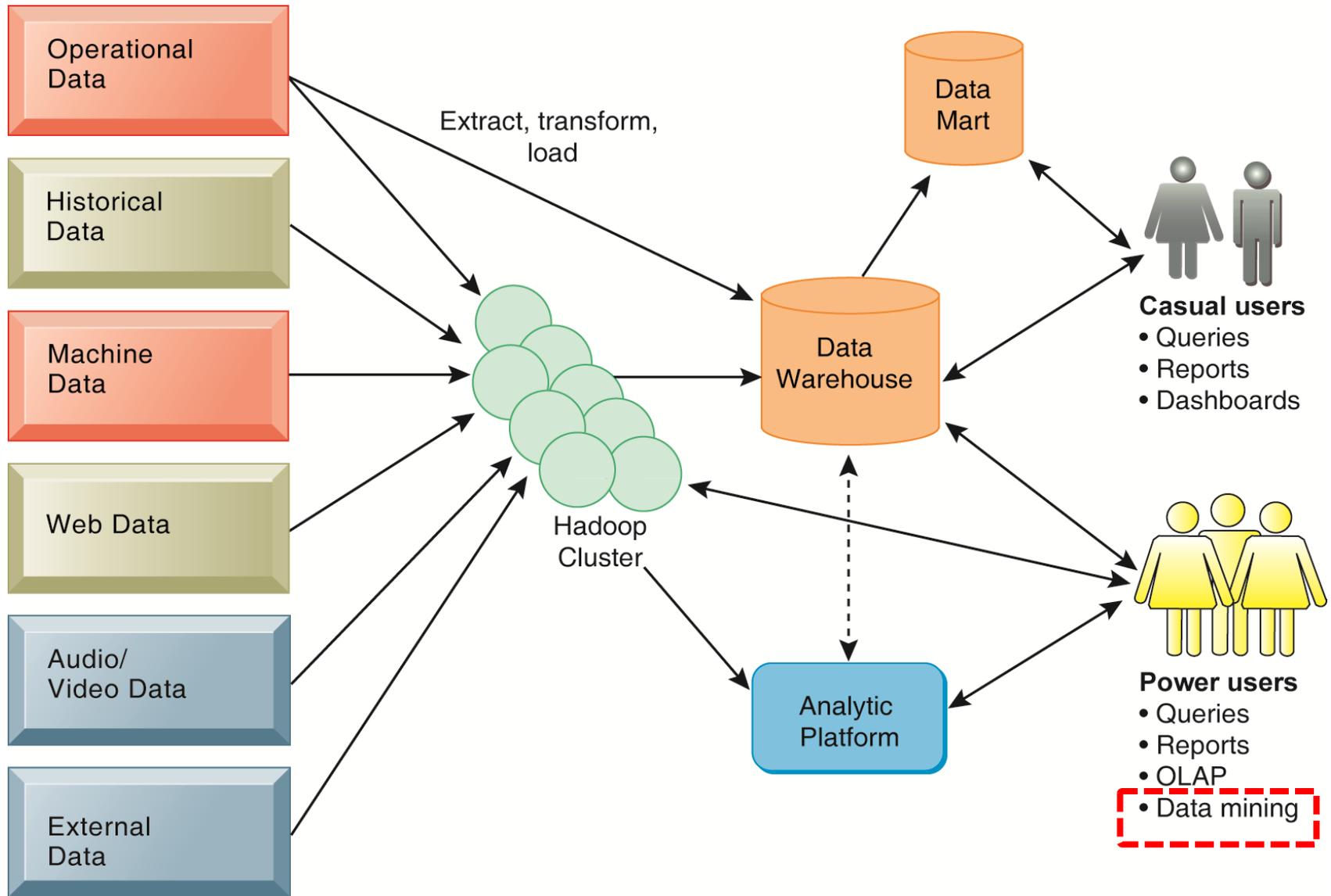
- Integrated analysis model
- Natural Language Processing
- Information Extraction
- Anomaly Detection
- Discovery of relationships among heterogeneous data
- Large-scale visualization
- Parallel distributed processing



Analysts

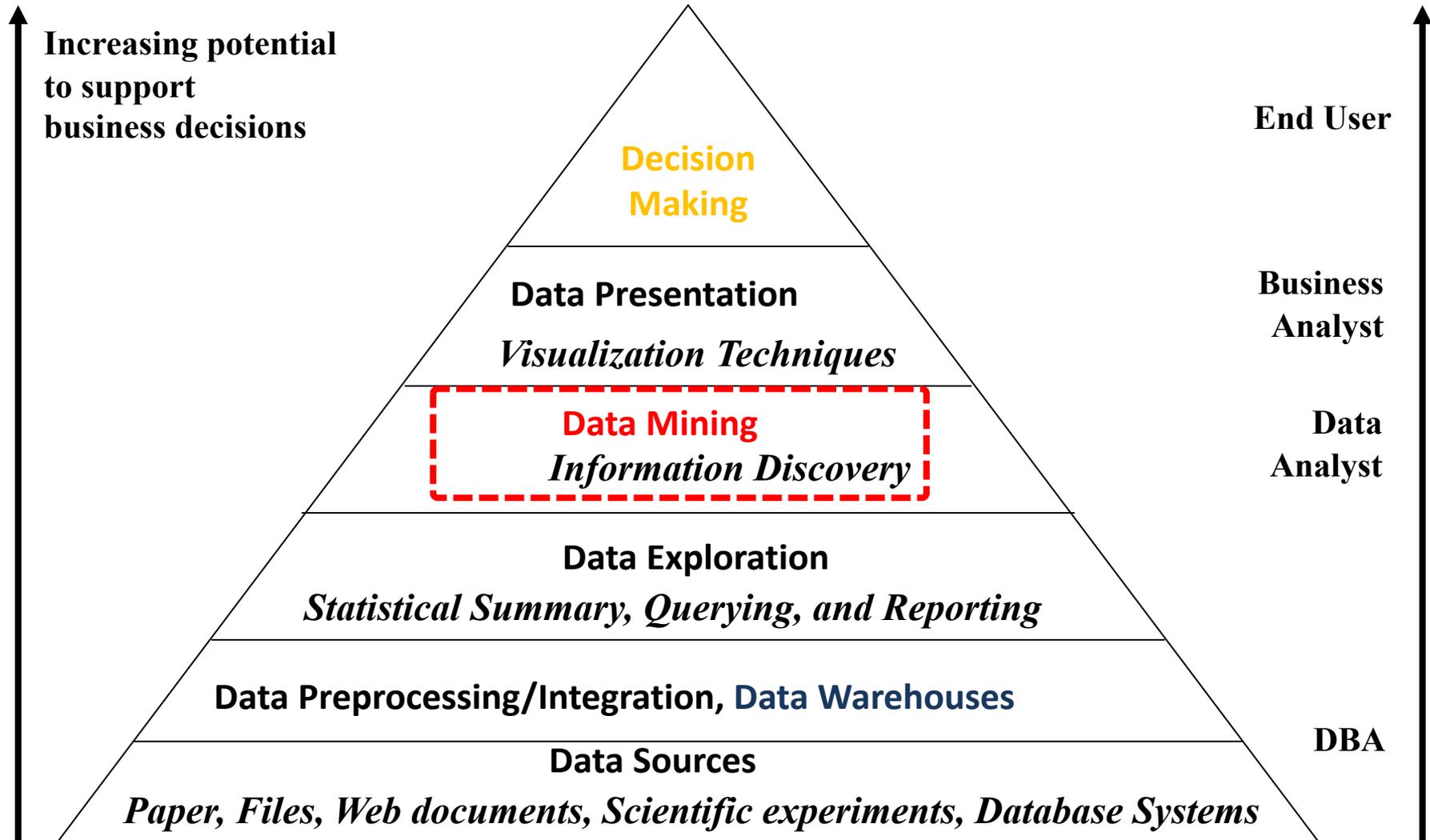
- Model Construction
- Explanation by Model
- Construction and confirmation of individual hypothesis
- Description and execution of application-specific task

Business Intelligence (BI) Infrastructure



Data Warehouse

Data Mining and Business Intelligence

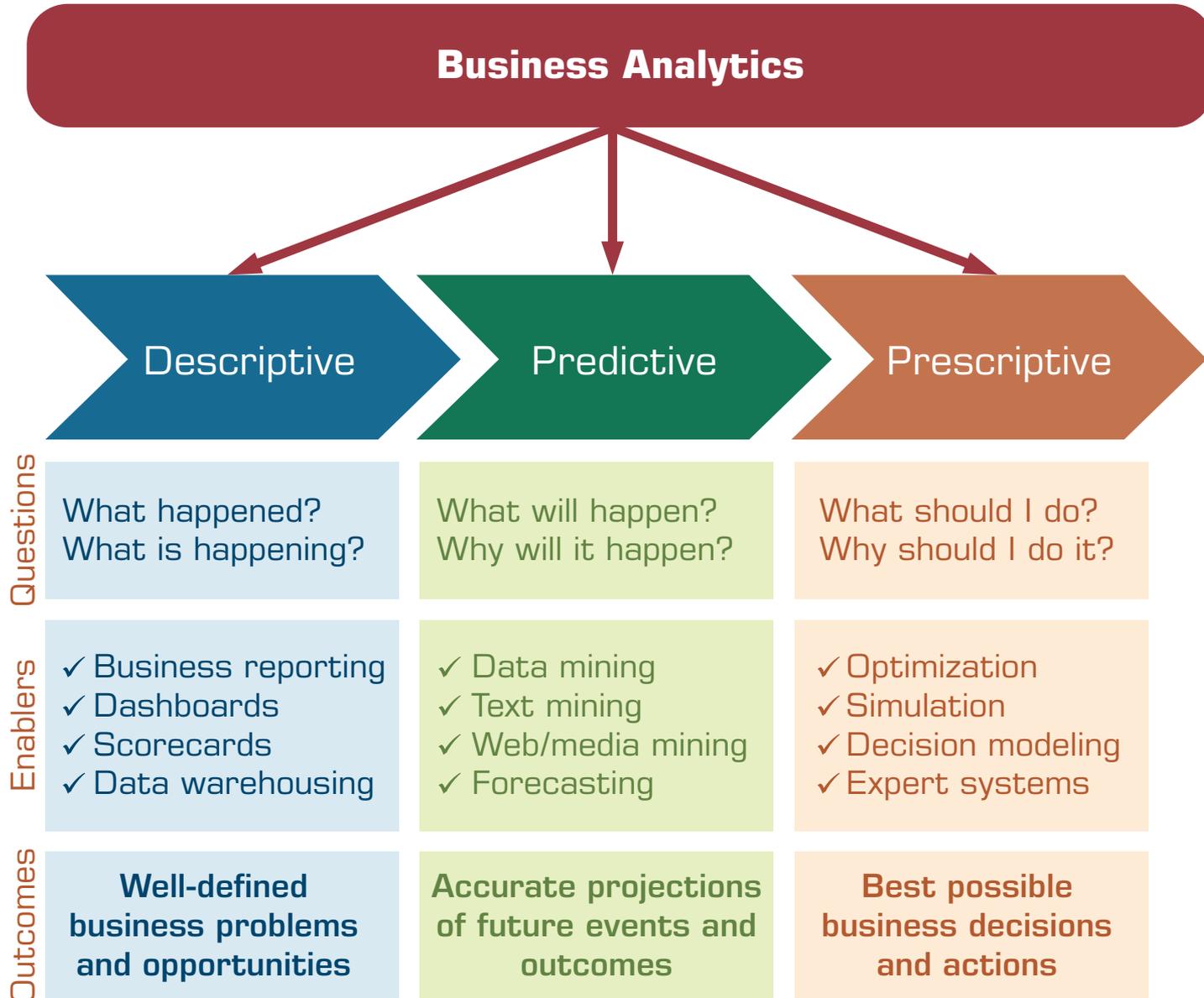


The Evolution of BI Capabilities

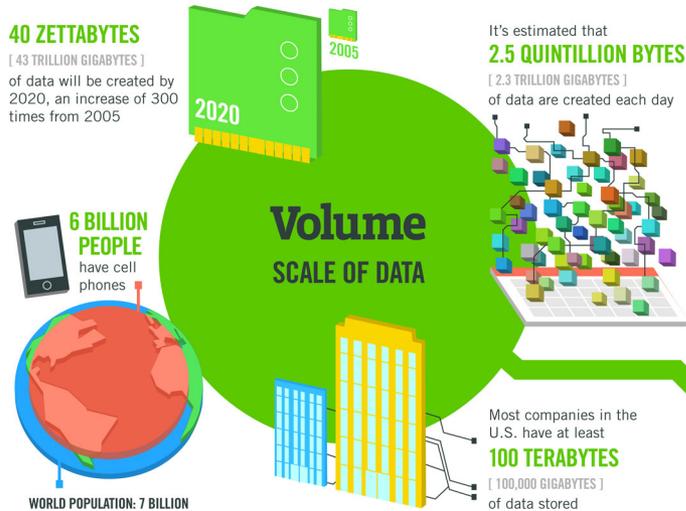


Source: Turban et al. (2011), Decision Support and Business Intelligence Systems

Three Types of Analytics



Big Data 4 V



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



As of 2011, the global size of data in healthcare was estimated to be

150 EXABYTES
[161 BILLION GIGABYTES]



30 BILLION PIECES OF CONTENT are shared on Facebook every month



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



400 MILLION TWEETS are sent per day by about 200 million monthly active users



Variety
DIFFERENT FORMS OF 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

Velocity
ANALYSIS OF STREAMING DATA

By 2016, it is projected there will be **18.9 BILLION NETWORK CONNECTIONS** – almost 2.5 connections per person on earth



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

Veracity
UNCERTAINTY OF DATA

in one survey were unsure of how much of their data was inaccurate

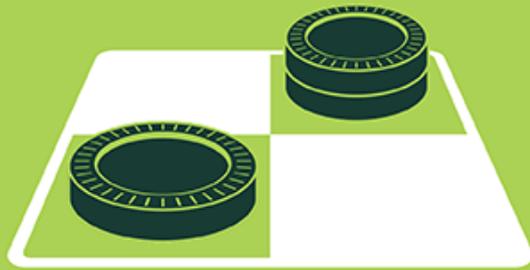
value

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.



1950's

1960's

1970's

1980's

1990's

2000's

2010's

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

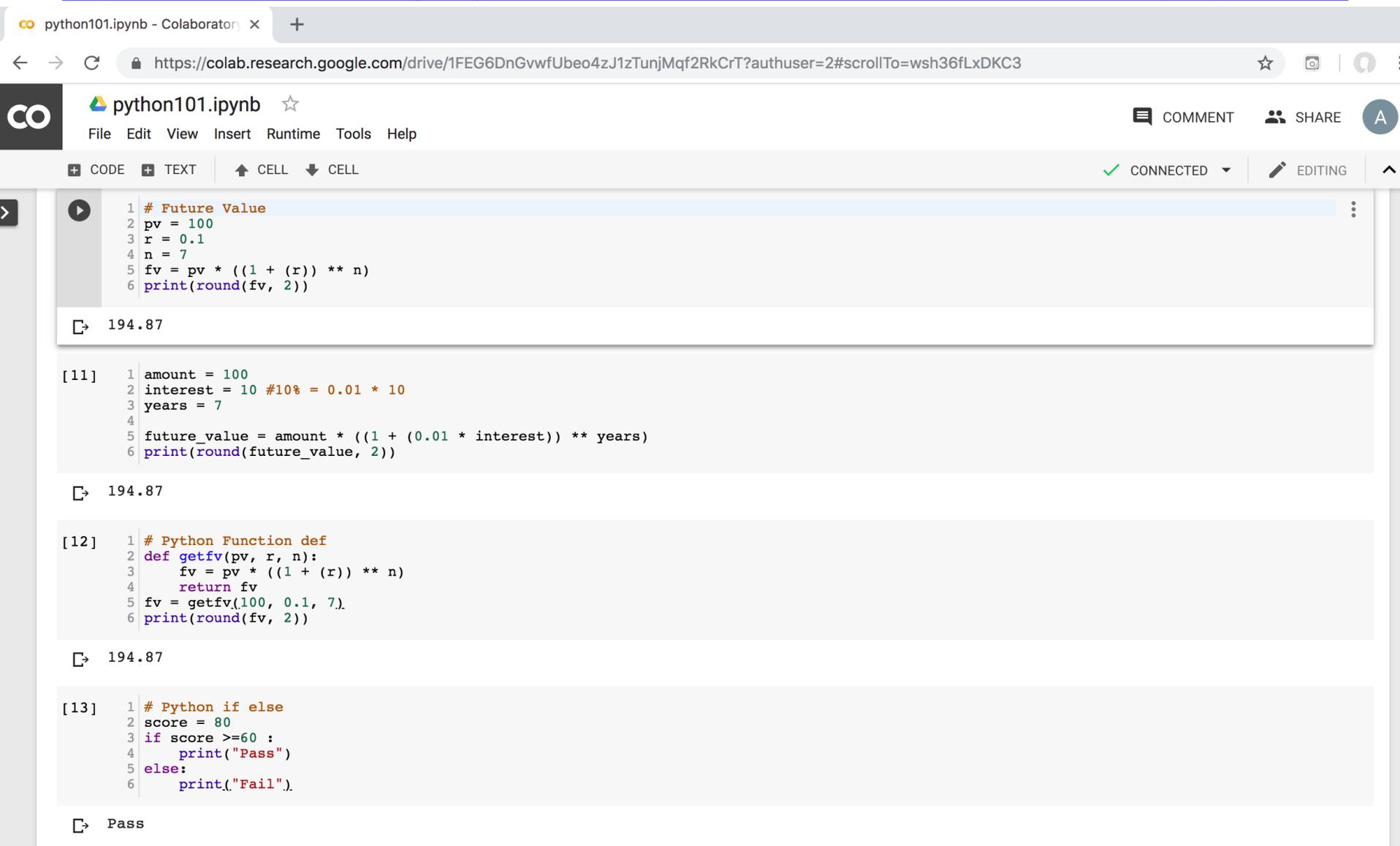
Reinforcement
Learning

Google Colab

The screenshot shows the Google Colab web interface. At the top, the browser address bar displays the URL <https://colab.research.google.com/notebooks/welcome.ipynb>. The main header includes the Colab logo, the text "Hello, Colaboratory", and a menu with options: File, Edit, View, Insert, Runtime, Tools, and Help. On the right side of the header, there is a "SHARE" button and a user profile picture. Below the header, a toolbar contains buttons for "+ CODE", "+ TEXT", "↑ CELL", "↓ CELL", and "COPY TO DRIVE". On the far right of the toolbar are "CONNECT" and "EDITING" options. A left-hand sidebar contains a "Table of contents" section with links to "Getting Started", "Highlighted Features", "TensorFlow execution", "GitHub", "Visualization", "Forms", "Examples", and "Local runtime support". The main content area features a large "Welcome to Colaboratory!" message with the Colab logo and a brief description: "Colaboratory is a free Jupyter notebook environment that requires no setup and runs entirely in the cloud. See our [FAQ](#) for more info." Below this is a "Getting Started" section with a list of links: "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", and "Machine Learning Crash Course: Intro to Pandas & First Steps with TensorFlow". A "Highlighted Features" section is partially visible, starting with a "Seedbank" subsection that says "Looking for Colab notebooks to learn from? Check out [Seedbank](#), a place to discover interactive machine learning examples." Below that, the "TensorFlow execution" subsection begins with the text "Colaboratory allows you to execute TensorFlow code in your browser with a single click. The example below adds two matrices." followed by a mathematical equation:
$$\begin{bmatrix} 1. & 1. & 1. \end{bmatrix} + \begin{bmatrix} 1. & 2. & 3. \end{bmatrix} = \begin{bmatrix} 2. & 3. & 4. \end{bmatrix}$$

Python in Google Colab (Python101)

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



python101.ipynb - Colaboratory

File Edit View Insert Runtime Tools Help

COMMENT SHARE

CONNECTED EDITING

```
1 # Future Value
2 pv = 100
3 r = 0.1
4 n = 7
5 fv = pv * ((1 + (r)) ** n)
6 print(round(fv, 2))
```

194.87

```
[11] 1 amount = 100
2 interest = 10 #10% = 0.01 * 10
3 years = 7
4
5 future_value = amount * ((1 + (0.01 * interest)) ** years)
6 print(round(future_value, 2))
```

194.87

```
[12] 1 # Python Function def
2 def getfv(pv, r, n):
3     fv = pv * ((1 + (r)) ** n)
4     return fv
5 fv = getfv(100, 0.1, 7)
6 print(round(fv, 2))
```

194.87

```
[13] 1 # Python if else
2 score = 80
3 if score >=60 :
4     print("Pass")
5 else:
6     print("Fail").
```

Pass

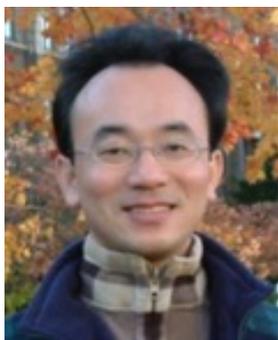
<https://tinyurl.com/aintpupython101>

Summary

- This course introduces the **fundamental concepts**, **research issues**, and **hands-on practices** of data mining.
- Topics include
 1. Introduction to data mining
 2. ABC: AI, Big Data, Cloud Computing
 3. Foundations of Data Mining in Python
 4. Data Science and Data Mining: Discovering, Analyzing, Visualizing and Presenting Data
 5. Unsupervised Learning: Association Analysis, Market Basket Analysis
 6. Unsupervised Learning: Cluster Analysis, Market Segmentation
 7. Supervised Learning: Classification and Prediction
 8. Machine Learning and Deep Learning
 9. Convolutional Neural Networks, Recurrent Neural Networks, Reinforcement Learning
 10. Social Network Analysis
 11. Case Study on Data Mining

資料探勘

(Data Mining)



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/>



aws academy

Accredited
Educator



aws
certified

Solutions
Architect
Associate



aws
certified

Cloud
Practitioner