



Big Data Mining

Course Orientation for Big Data Mining

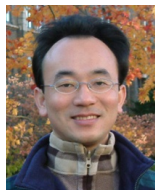
1071BDM01

TLVXM1A (M2244) (8619) (Fall 2018)

(MBA, DBETKU) (3 Credits, Required) [Full English Course]

(Master's Program in Digital Business and Economics)

Mon, 9, 10, 11, (16:10-19:00) (B206)



Min-Yuh Day, Ph.D.

Assistant Professor

Department of Information Management

Tamkang University

<http://mail.tku.edu.tw/myday>

2018-09-10





Course Syllabus

Tamkang University

Academic Year 107, 1st Semester (Fall, 2018)

- Course Title: **Big Data Mining**
- Instructor: Min-Yuh Day
- Course Class: TLVXM1A (MBA DBETKU)
 - Master's Program in Digital Business and Economics, 1A
- Details
 - Required
 - One Semester
 - 3 Credits
- Time & Place: Mon, 9, 10, 11, (16:10-19:00) (B206)



Department Teaching Objectives

- Train students not only to acquire knowledge from economics, finance, and industrial developments but also to apply information technology and analytical and quantitative skills to various situations.
- Students can enhance their competitiveness in facing rapid changes in world economy.



Department Core Competences

1. Cultivating students the ability of computer programming.
2. Training students the ability of website design for starting up a business.
3. Training students the ability of analyzing various situations in the financial market.
4. Helping students to acquire the knowledge of financial technology.

Course Introduction



- This course introduces the **fundamental concepts** and **research issues** of **Big Data Mining**.
- Topics include
 - ABC: AI, Big Data, Cloud Computing,
 - Data Science and Big Data Analytics: Discovering, Analyzing, Visualizing and Presenting Data,
 - Fundamental Big Data: MapReduce Paradigm, Hadoop and Spark Ecosystem,
 - Foundations of Big Data Mining in Python,
 - Supervised Learning: Classification and Prediction,
 - Unsupervised Learning: Cluster Analysis,
 - Unsupervised Learning: Association Analysis,
 - Machine Learning with Scikit-Learn in Python,
 - Deep Learning for Finance Big Data with TensorFlow,
 - Convolutional Neural Networks (CNN)
 - Recurrent Neural Networks (RNN)
 - Reinforcement Learning (RL)
 - Social Network Analysis (SNA)



Teaching Objectives

1. Understand and apply the fundamental concepts and research issues of big data mining.
2. Conduct information systems research in the context of big data mining.



Teaching Methods

- Lecture
- Discussion
- Simulation
- Practicum
- Problem Solving



Assessment

- Practicum
 - Report
- Participation

Course Schedule (1/2)



Week	Date	Subject/Topics
1	2018/09/10	Course Orientation for Big Data Mining
2	2018/09/17	ABC: AI, Big Data, Cloud Computing
3	2018/09/24	Mid-Autumn Festival (Day off)
4	2018/10/01	Data Science and Big Data Analytics: Discovering, Analyzing, Visualizing and Presenting Data
5	2018/10/08	Fundamental Big Data: MapReduce Paradigm, Hadoop and Spark Ecosystem
6	2018/10/15	Foundations of Big Data Mining in Python
7	2018/10/22	Supervised Learning: Classification and Prediction
8	2018/10/29	Unsupervised Learning: Cluster Analysis
9	2018/11/05	Unsupervised Learning: Association Analysis

Course Schedule (2/2)



Week	Date	Subject/Topics
10	2018/11/12	Midterm Project Report
11	2018/11/19	Machine Learning with Scikit-Learn in Python
12	2018/11/26	Deep Learning for Finance Big Data with TensorFlow
13	2018/12/03	Convolutional Neural Networks (CNN)
14	2018/12/10	Recurrent Neural Networks (RNN)
15	2018/12/17	Reinforcement Learning (RL)
16	2018/12/24	Social Network Analysis (SNA)
17	2018/12/31	Bridge Holiday (Extra Day Off)
18	2019/01/07	Final Project Presentation



Grading Policy

- Mark of Usual: 40%
- Final Project: 60%
 - Midterm Project Report
 - Final Project Report



Textbooks and References

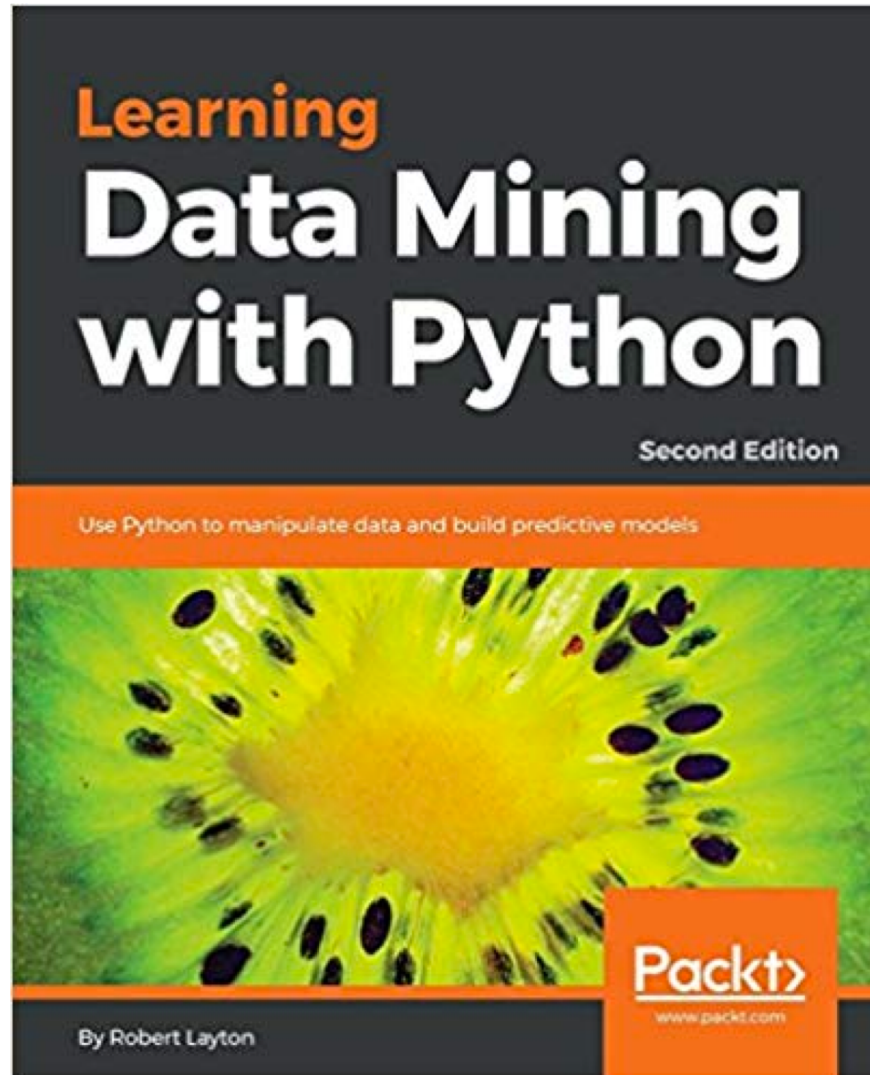
- Textbook: Slides

- <http://mail.tku.edu.tw/myday/teaching.htm#1071BDM>

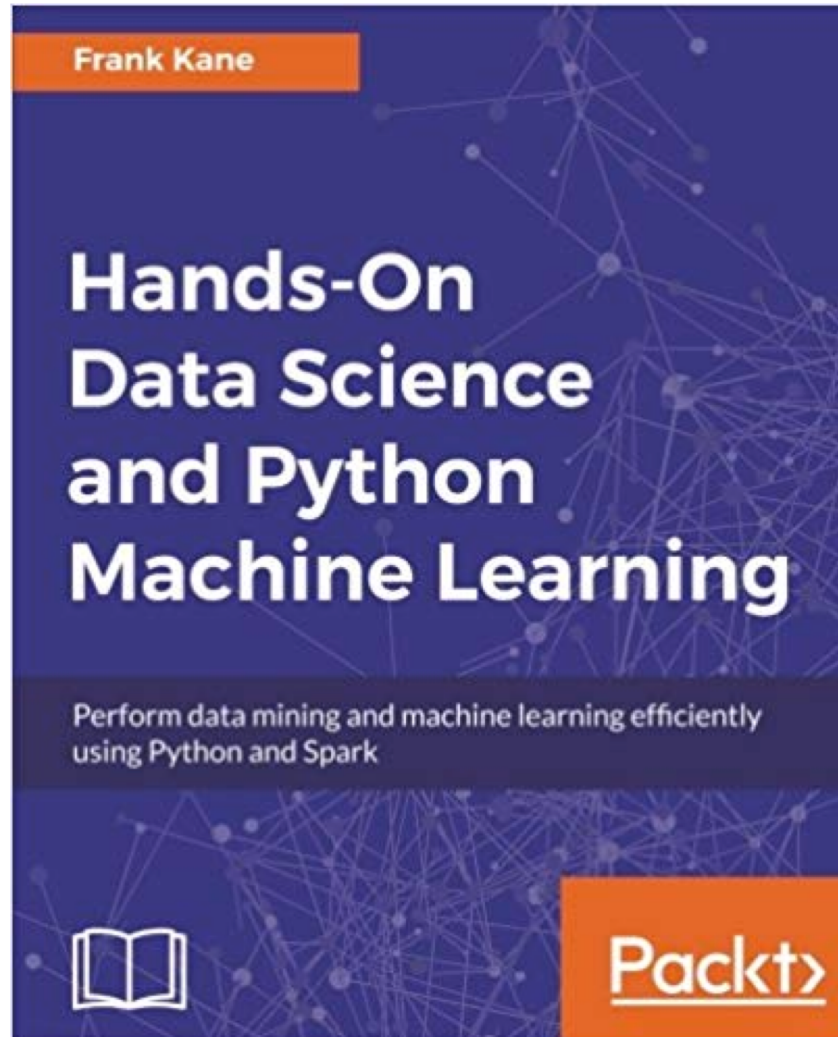
- References

- Learning Data Mining with Python - Second Edition, Robert Layton, Packt Publishing, 2017.
 - Hands-On Data Science and Python Machine Learning: Perform data mining and machine learning efficiently using Python and Spark, Frank Kane, Packt Publishing, 2017.
 - Hands-On Machine Learning with Scikit-Learn and TensorFlow: Concepts, Tools, and Techniques to Build Intelligent Systems, Aurélien Géron, O'Reilly Media, 2017
 - Practical Machine Learning with Python: A Problem-Solver's Guide to Building Real-World Intelligent Systems, Dipanjan Sarkar, Raghav Bali, Tushar Sharma, Apress, 2017.
 - Python Machine Learning: Machine Learning and Deep Learning with Python, scikit-learn, and TensorFlow, 2nd Edition, Sebastian Raschka and Vahid Mirjalili, Packt Publishing, 2017.
 - Deep Learning with Python, Francois Chollet, Manning Publications, 2017.
 - Big Data, Data Mining, and Machine Learning: Value Creation for Business Leaders and Practitioners, Jared Dean, Wiley, 2014.
 - Data Mining: Concepts and Techniques, Third Edition, Jiawei Han, Micheline Kamber and Jian Pei, Morgan Kaufmann, 2011.

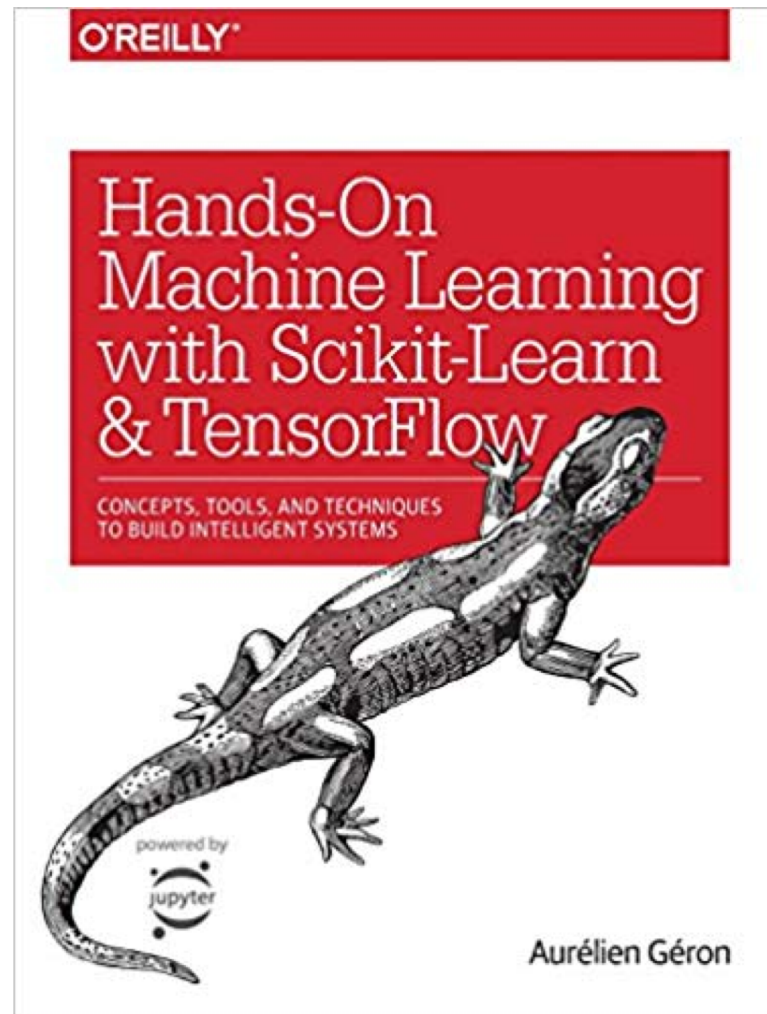
Learning Data Mining with Python - Second Edition,
Robert Layton,
Packt Publishing, 2017



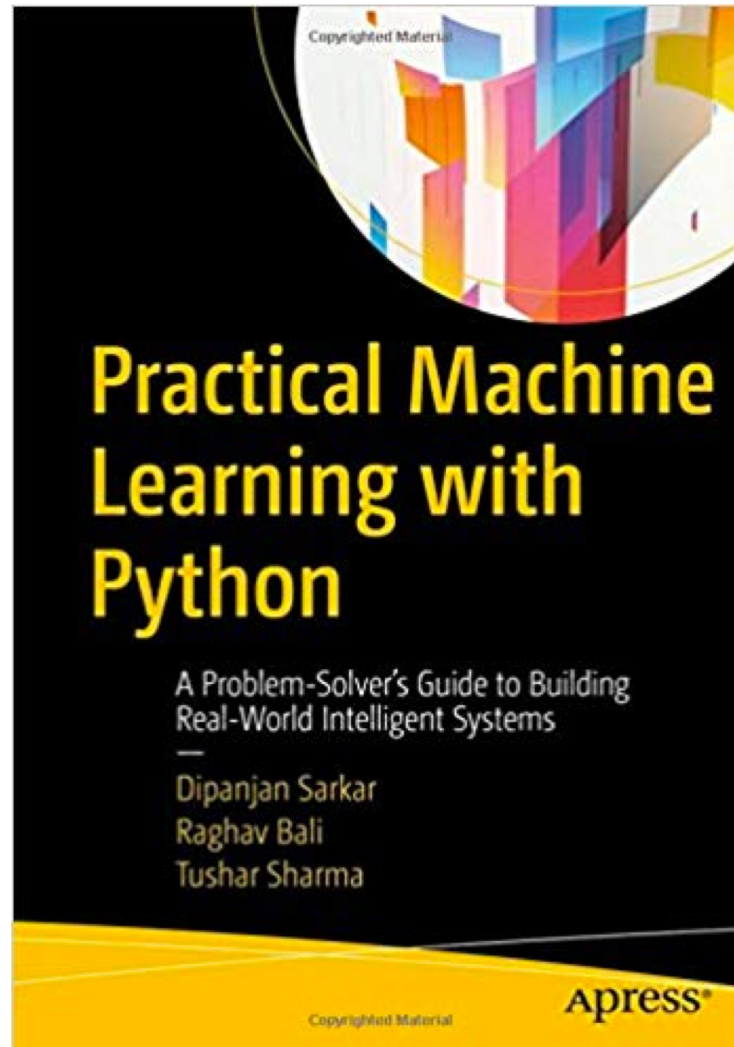
Hands-On Data Science and Python Machine Learning: Perform data mining and machine learning efficiently using Python and Spark, Frank Kane, Packt Publishing, 2017



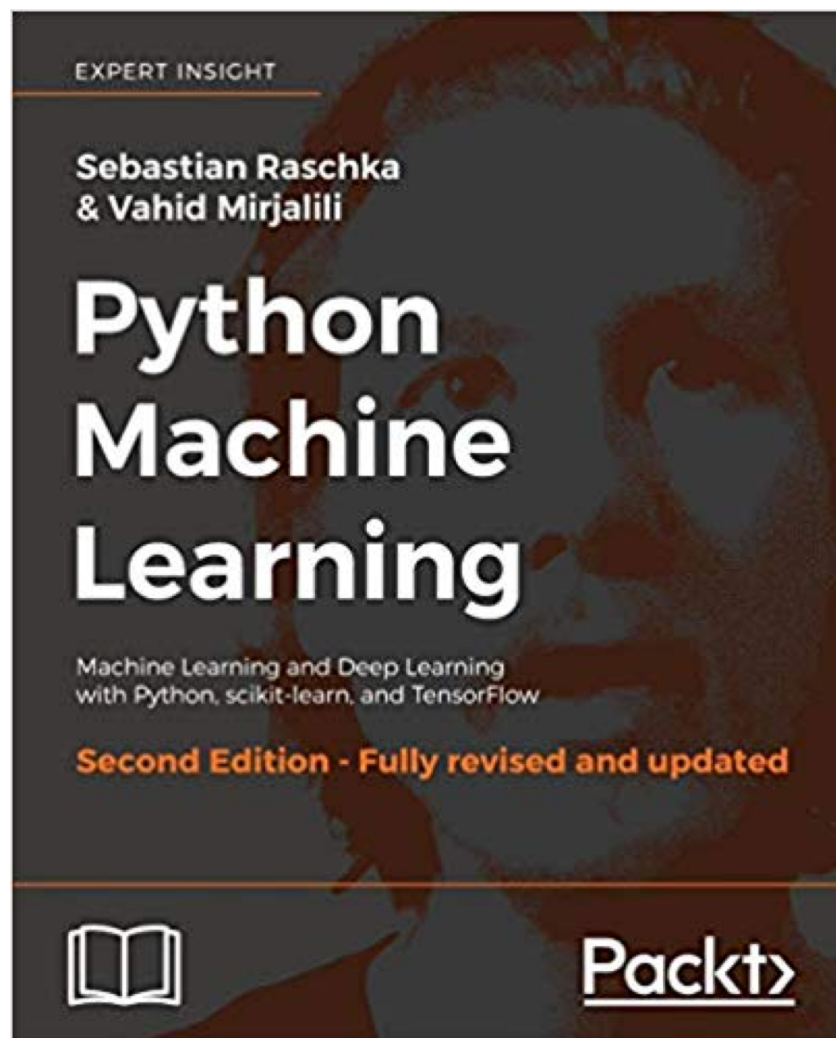
Hands-On Machine Learning with Scikit-Learn and TensorFlow: Concepts, Tools, and Techniques to Build Intelligent Systems, Aurélien Géron, O'Reilly Media, 2017



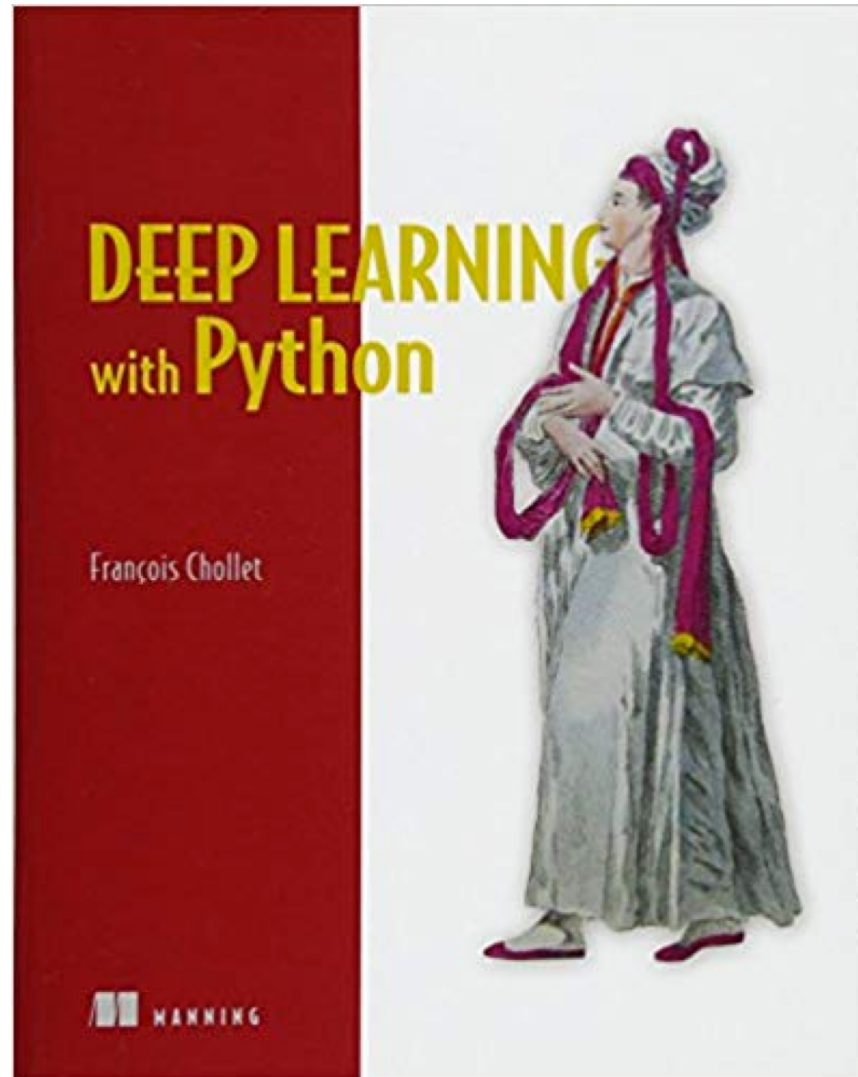
Practical Machine Learning with Python: A Problem-Solver's Guide to Building Real-World Intelligent Systems, Dipanjan Sarkar, Raghav Bali, Tushar Sharma, Apress, 2017.



**Python Machine Learning: Machine Learning and Deep Learning
with Python, scikit-learn, and TensorFlow, 2nd Edition,
Sebastian Raschka and Vahid Mirjalili,
Packt Publishing, 2017.**

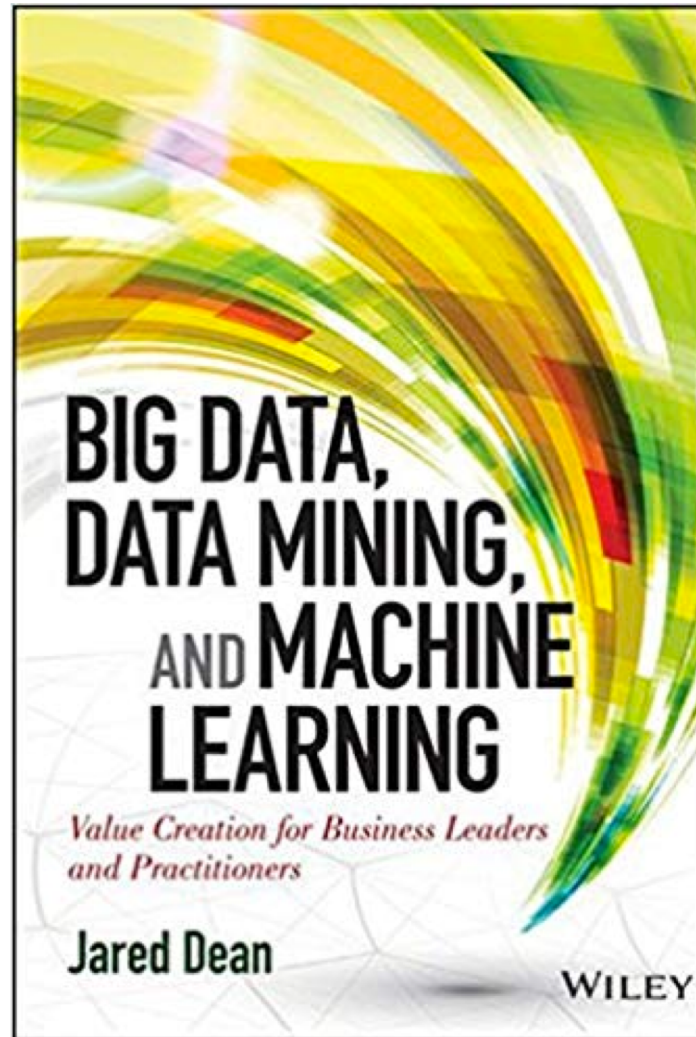


Deep Learning with Python,
Francois Chollet,
Manning Publications, 2017.

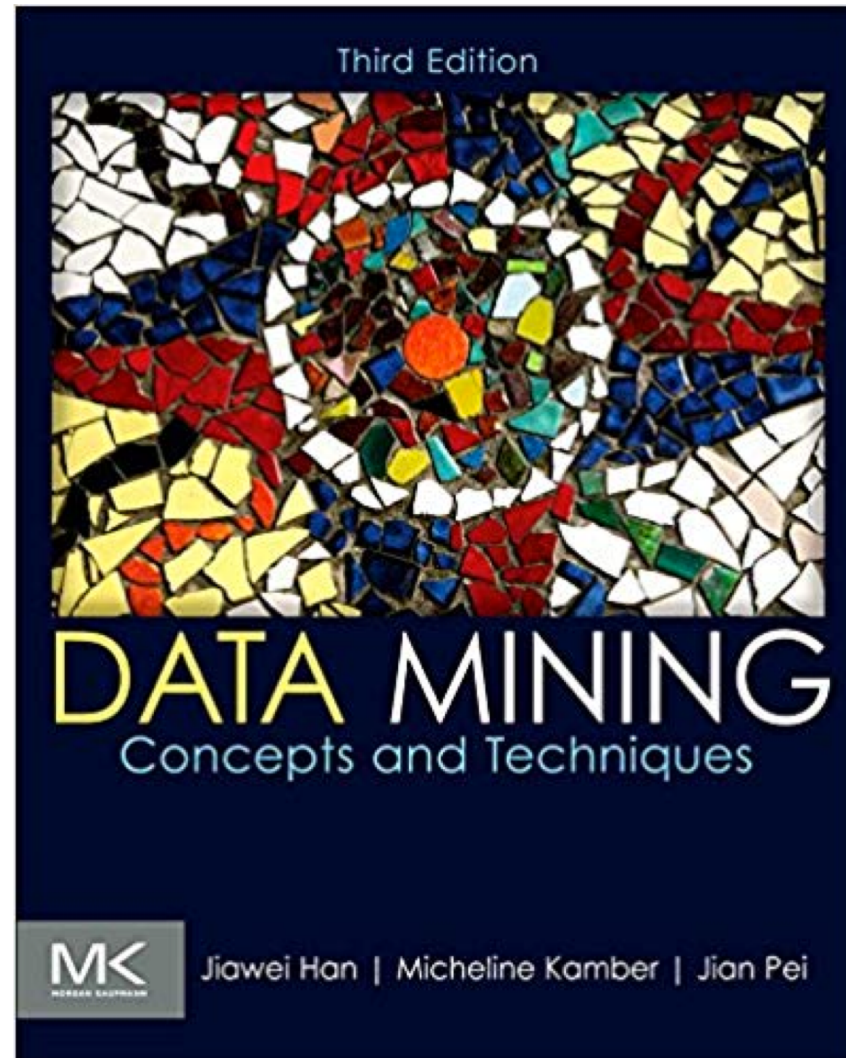


Source: <https://www.amazon.com/Deep-Learning-Python-Francois-Chollet/dp/1617294438>

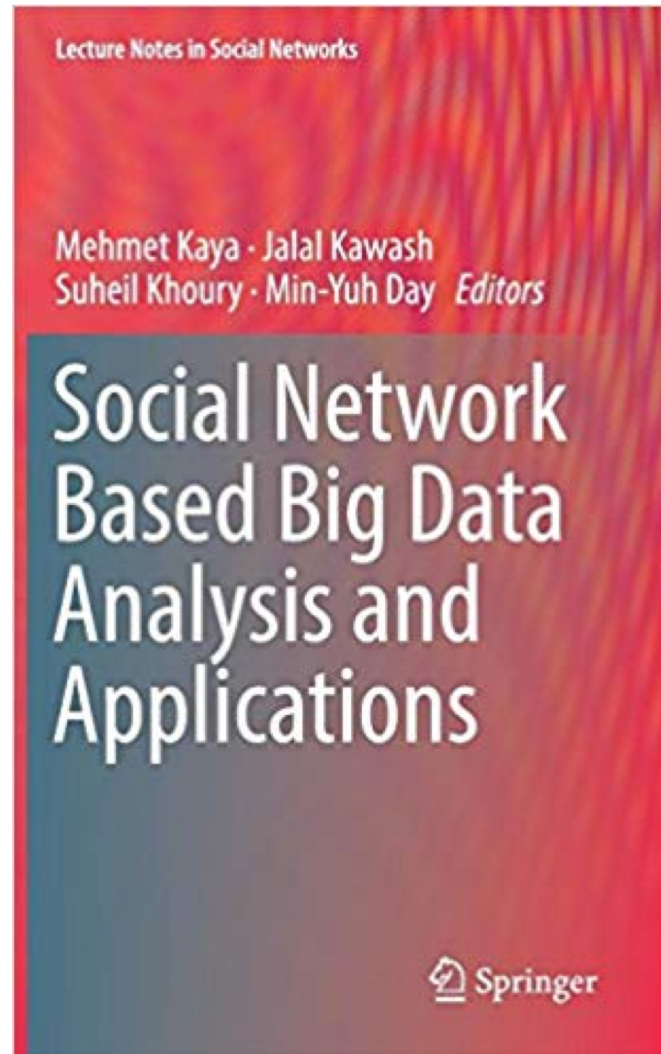
**Big Data, Data Mining, and Machine Learning: Value Creation for
Business Leaders and Practitioners,
Jared Dean,
Wiley, 2014.**



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



**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.**



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" (with up and down arrows), and "COPY TO DRIVE". To the right of the toolbar are "CONNECT" and "EDITING" options. A left-hand navigation sidebar is visible, containing 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, there 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:
$$\begin{bmatrix} 1. & 1. & 1. \end{bmatrix} + \begin{bmatrix} 1. & 2. & 3. \end{bmatrix} = \begin{bmatrix} 2. & 3. & 4. \end{bmatrix}$$

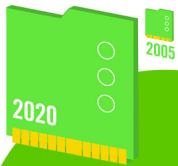
Big Data
Analytics
and
Data Mining

Big Data 4 V

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



Volume SCALE OF DATA



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**
[100,000 GIGABYTES]
of data stored



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]

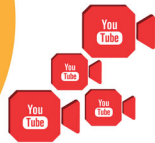


30 BILLION PIECES OF CONTENT are shared on Facebook every month



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



Velocity ANALYSIS OF STREAMING DATA



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



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

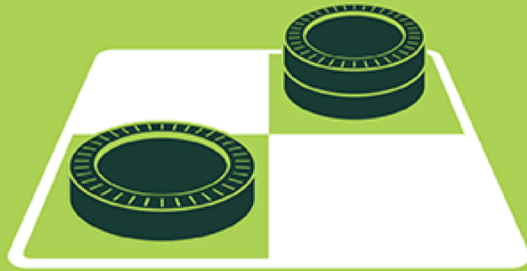
value

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

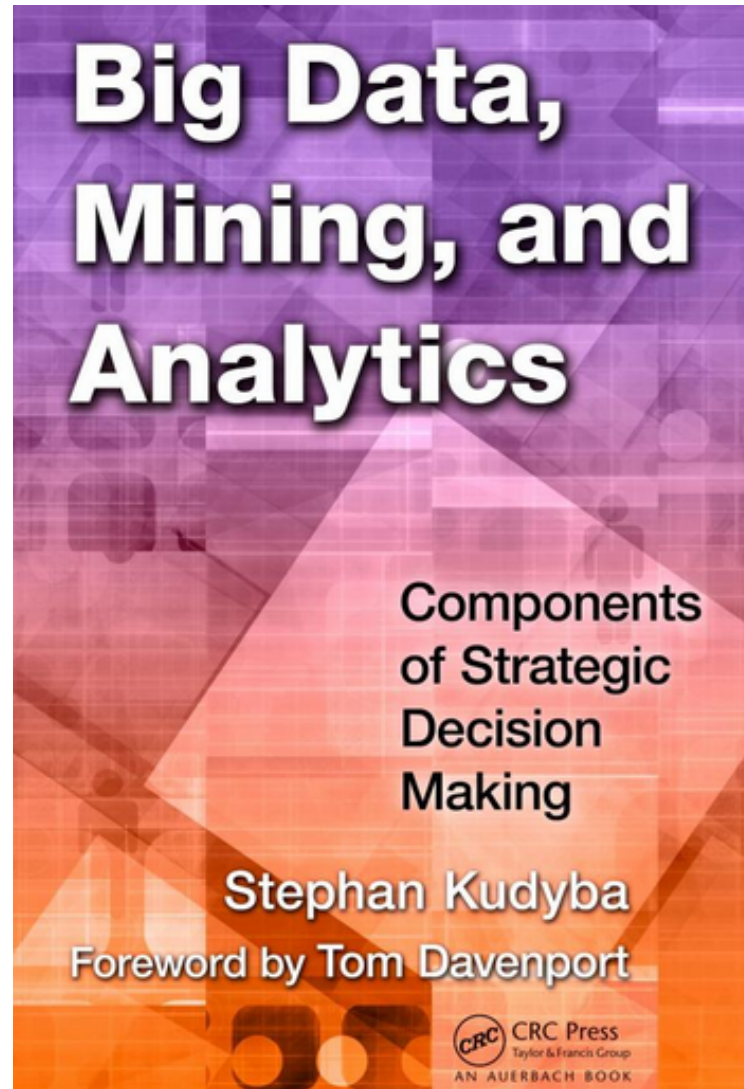
1990's

2000's

2010's

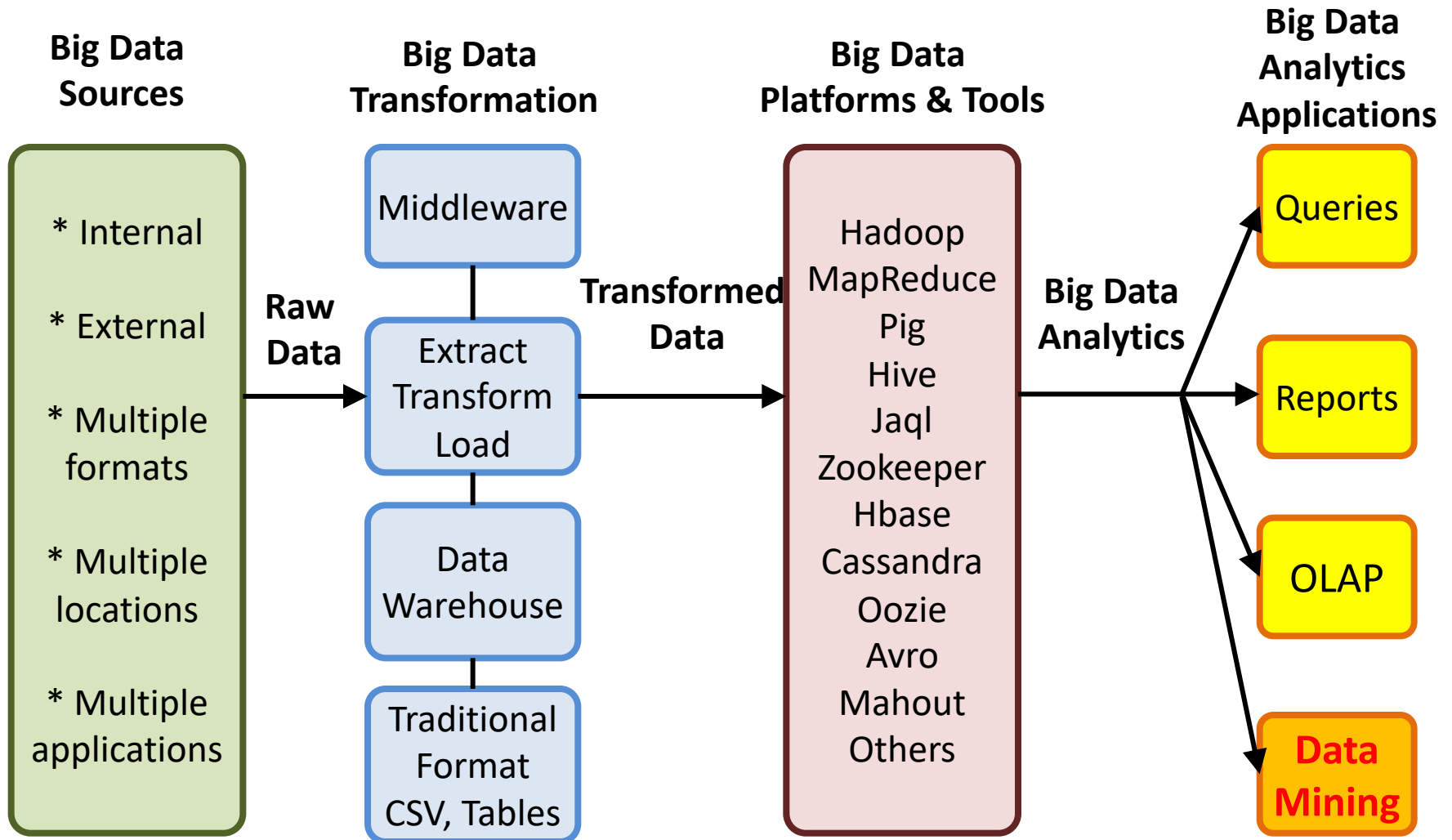
Since an early flush of optimism in the 1950s, smaller subsets of artificial intelligence – first machine learning, then deep learning, a subset of machine learning – have created ever larger disruptions.

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

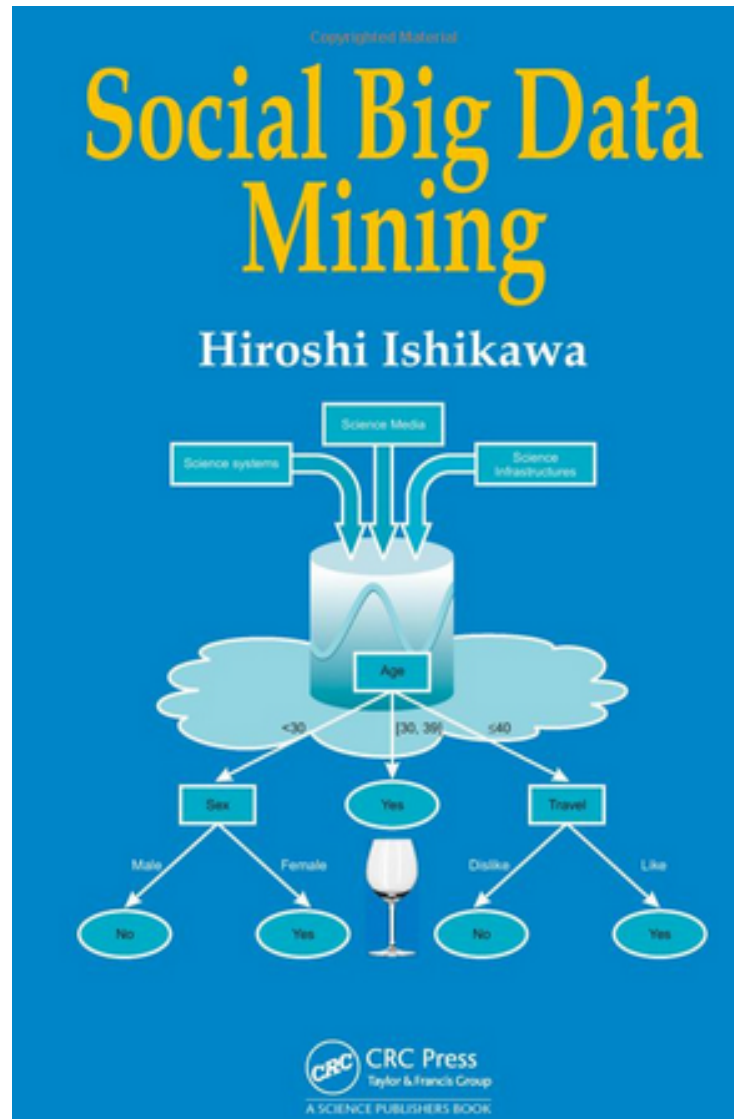


Architecture of Big Data Analytics



Social Big Data Mining

(Hiroshi Ishikawa, 2015)

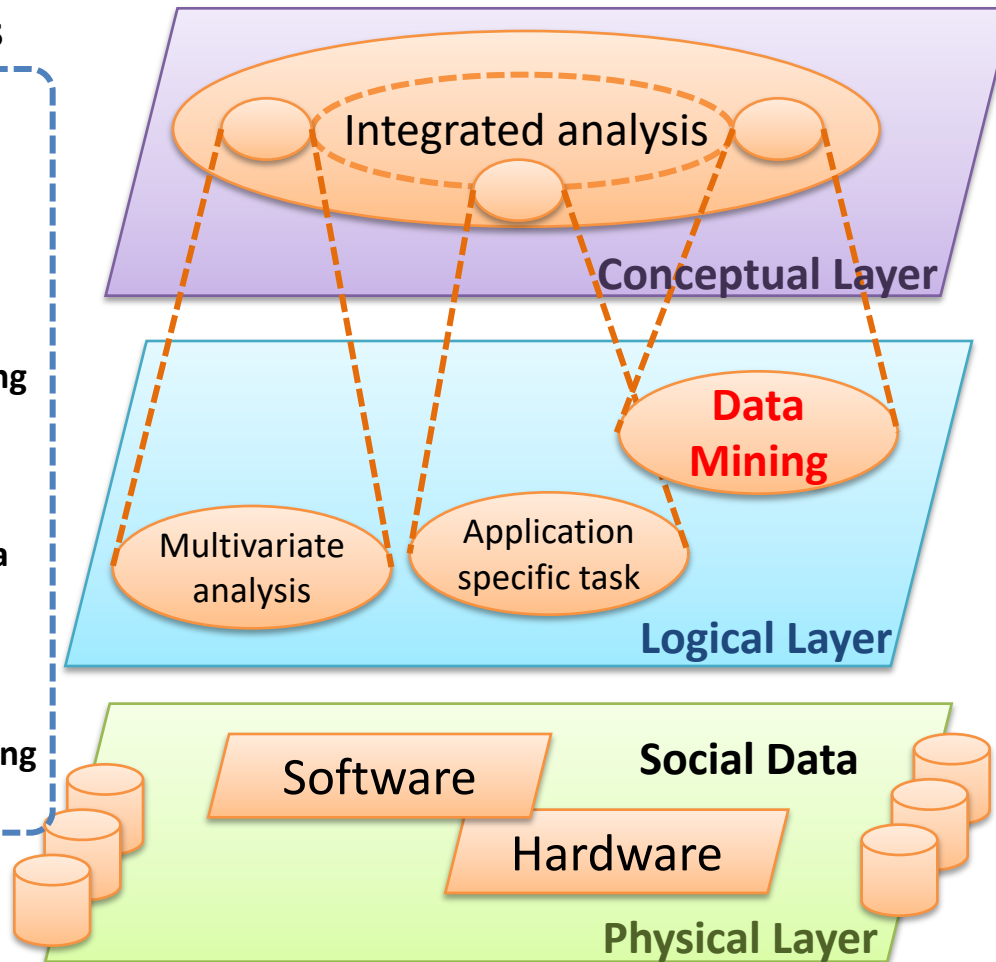


Architecture for Social Big Data Mining

(Hiroshi Ishikawa, 2015)

Enabling Technologies

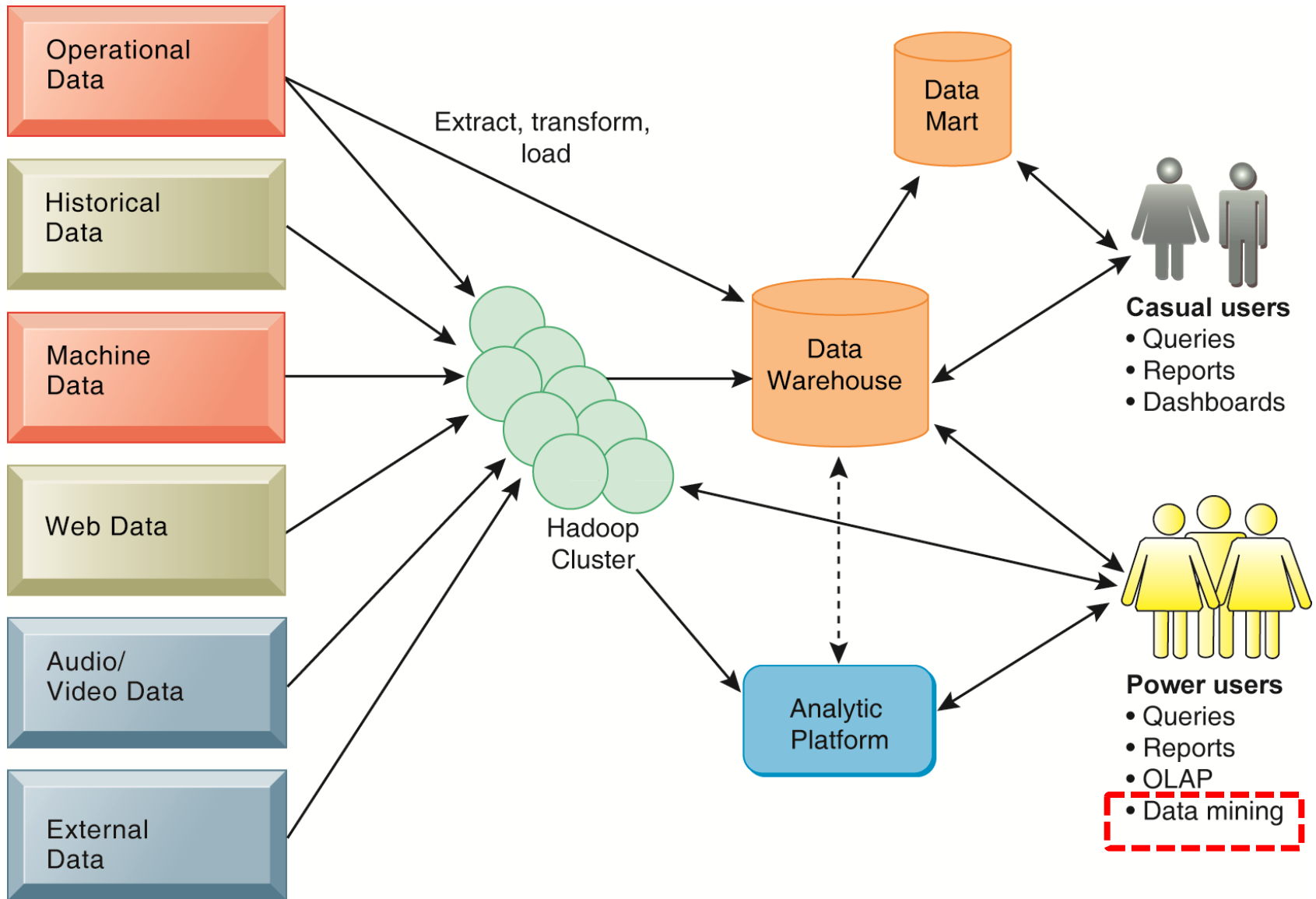
- Integrated analysis model
- Natural Language Processing
- Information Extraction
- Anomaly Detection
- Discovery of relationships among heterogeneous data
- Large-scale visualization
- Parallel distributed processing



Analysts

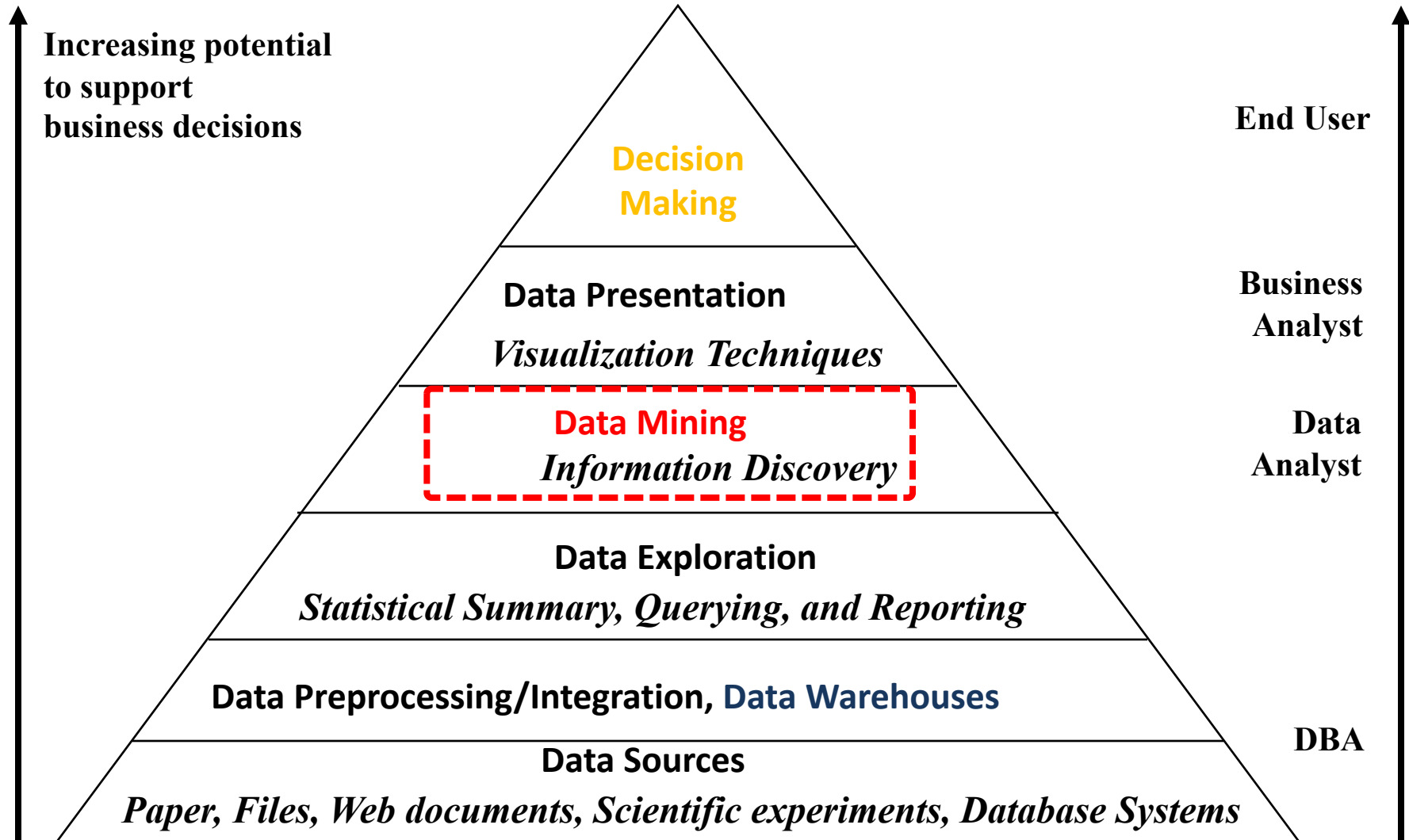
- Model Construction
- Explanation by Model
- Construction and confirmation of individual hypothesis
- Description and execution of application-specific task

Business Intelligence (BI) Infrastructure

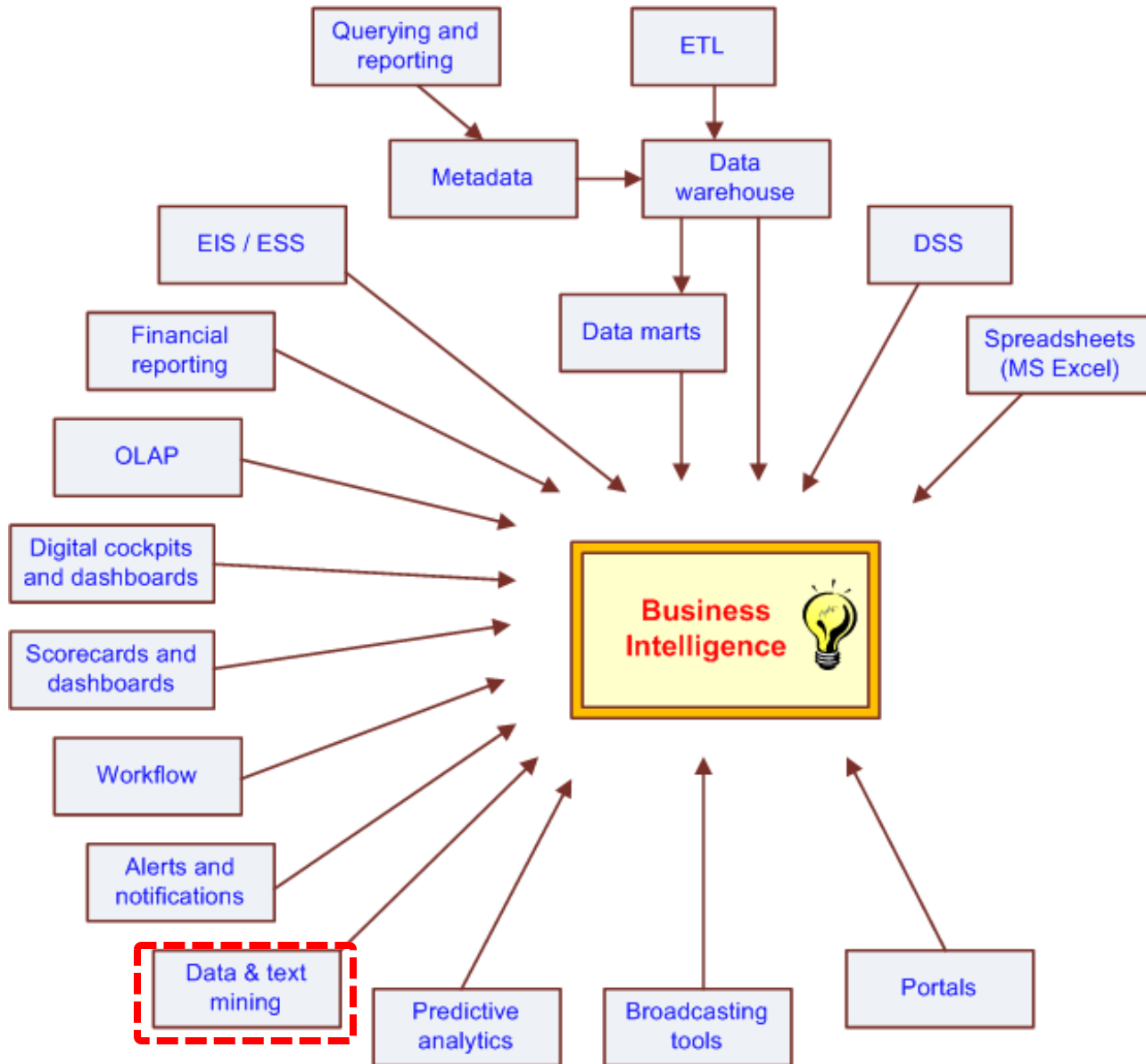


Data Warehouse

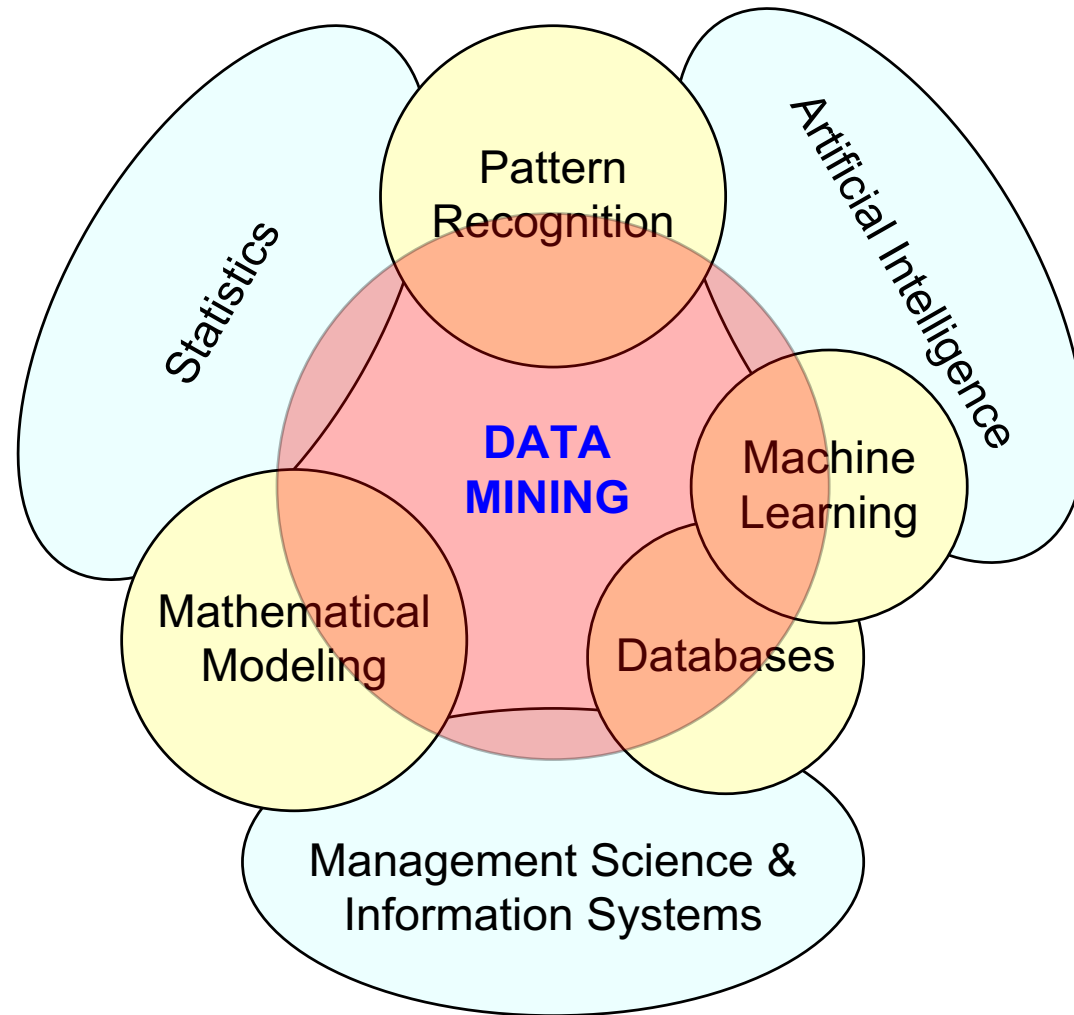
Data Mining and Business Intelligence



The Evolution of BI Capabilities



Data Mining at the Intersection of Many Disciplines





Data Mining:

Core **Analytics** Process

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

Fayyad, U., Piatetsky-Shapiro, G., & Smyth, P. (1996).

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,

Usama Fayyad,
Gregory Piatetsky-Shapiro,
and Padhraic Smyth

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

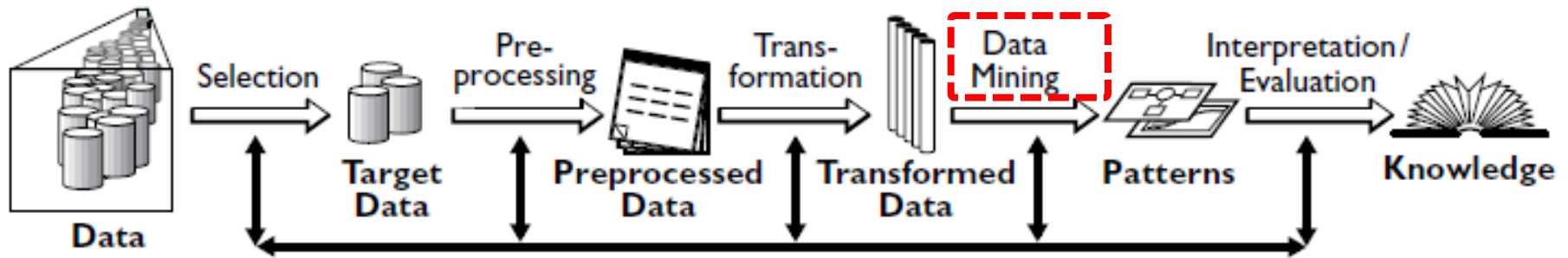


TEHRAN UNIVERSITY

Data Mining

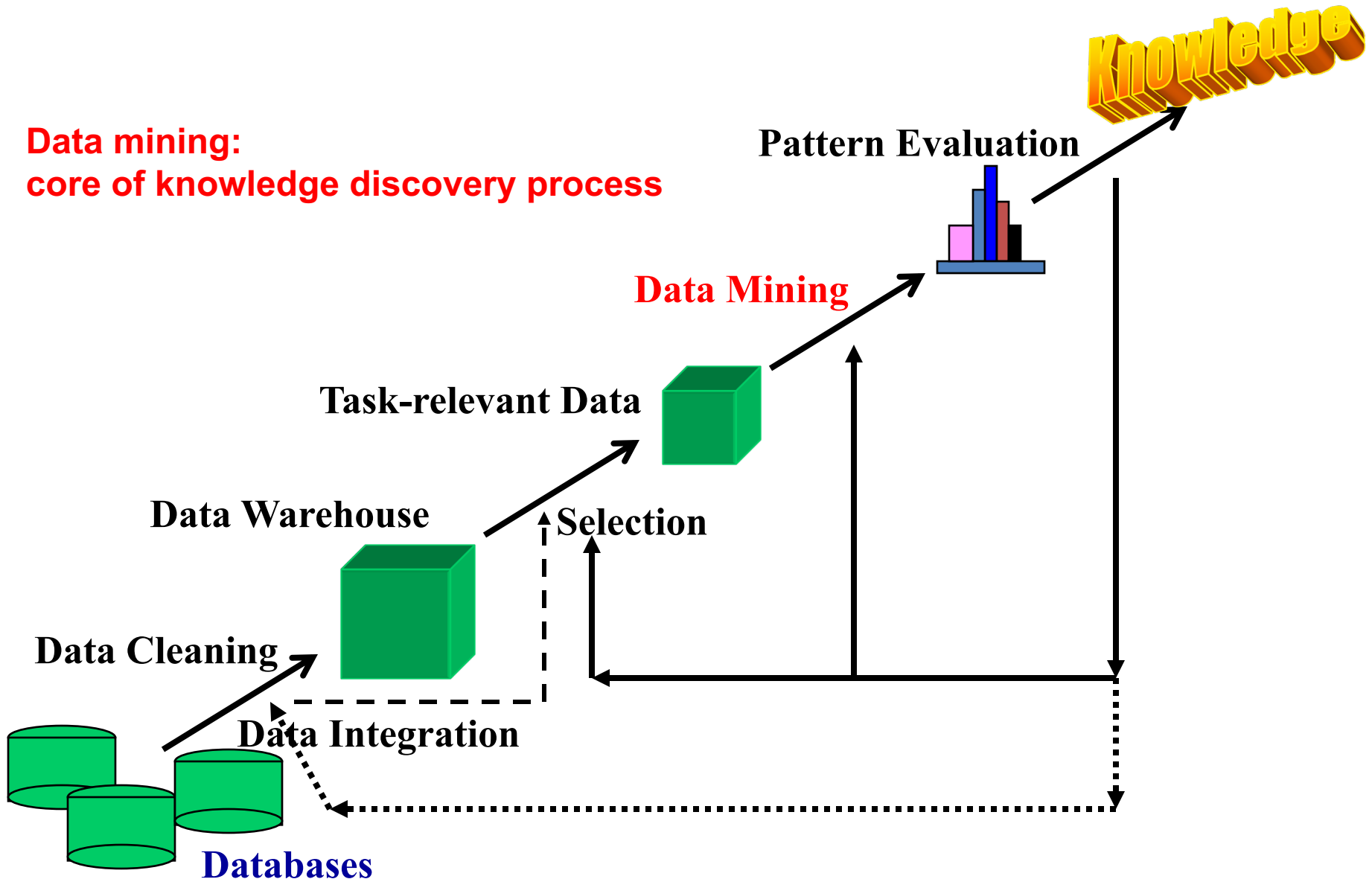
Knowledge Discovery in Databases (KDD) Process

(Fayyad et al., 1996)



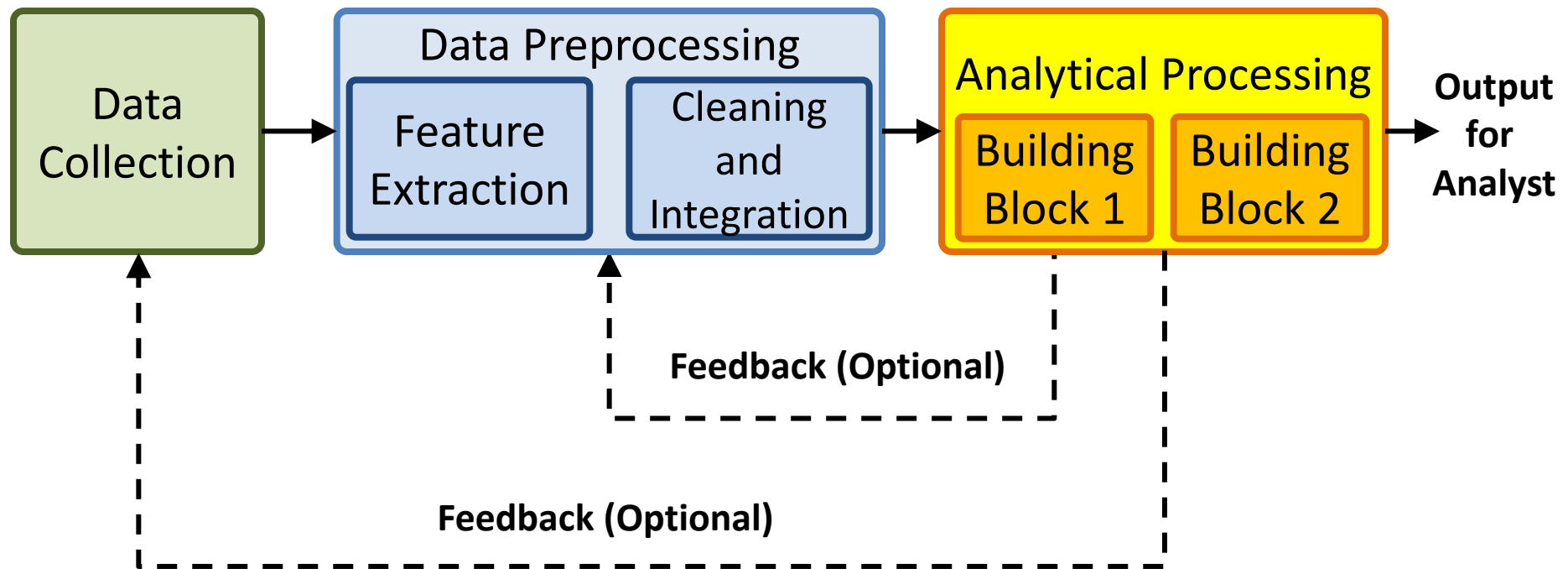
Knowledge Discovery (KDD) Process

Data mining:
core of knowledge discovery process

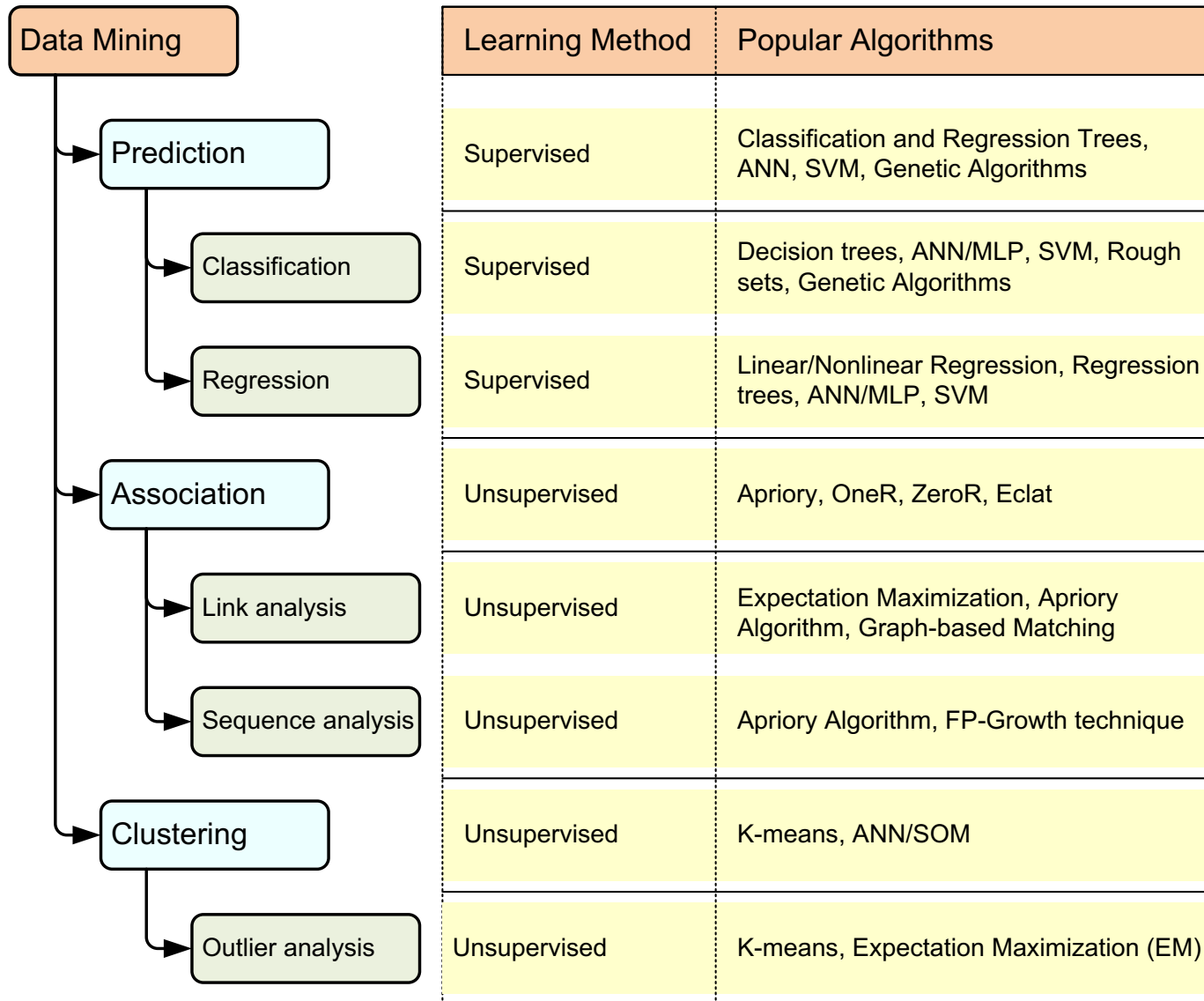


Data Mining Processing Pipeline

(Charu Aggarwal, 2015)



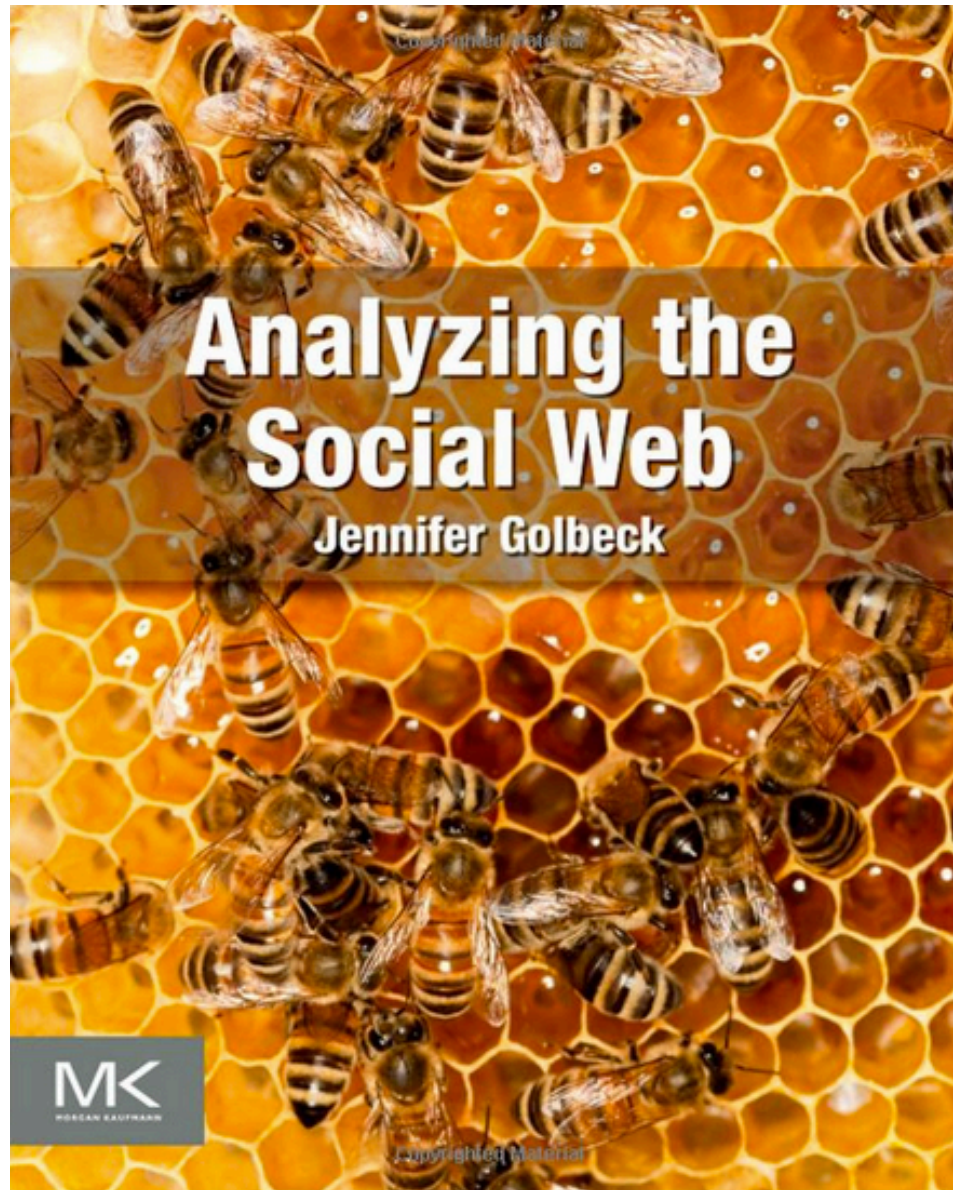
A Taxonomy for Data Mining Tasks



Business Insights with Social Analytics

Analyzing the Social Web: Social Network Analysis

Jennifer Golbeck (2013), *Analyzing the Social Web*, Morgan Kaufmann



The 14th NTCIR (2018 - 2019)

NTCIR (NII Testbeds and Community for Information access Research) Project



Japanese



About NTCIR



FAQ

Search



Publications/
Online Proceedings

Data/Tools

NTCIR CMS Site

Related URL's

Contact us

NTCIR Home > NTCIR-14

NTCIR 14

NTCIR-14 Conference

NEWS

NTCIR-14 Aims

Call for Task Proposals

How to Participate

Task Participation

Task Overview/Call for
Task Participation

User Agreement Forms

Organization

Important Dates

Contact Us

NTCIR 13

NTCIR 12

NTCIR-14

The 14th NTCIR (2018 - 2019)

Evaluation of Information Access Technologies

January 2018 - June 2019

What's New

NEW February 1, 2018: [Call for participation to the NTCIR-14 Kick-Off Event released.](#)

NEW February 1, 2018: Call for participation to the NTCIR-14 QALab-PoliInfo Kick-Off Event released.

December 5, 2017: The NTCIR-14 Task Selection Committee has selected the following six Tasks. Lifelig-3, OpenLiveQ-2, QA Lab-4, STC-3, WWW-2, CENTRE.

August 23, 2017: [NTCIR-14 Call for Task Proposals released.](#)(Closed.)

NEW About Proceedings

After the NTCIR-14 conference, a post-proceedings of revised selected papers will be published in [the Springer Lecture Notes on Computer Science \(LNCS\) series.](#)

<http://research.nii.ac.jp/ntcir/ntcir-14/index.html>

Lecture Notes in
Computer Science

NTCIR-14

Short Text Conversation Task (STC-3)

NTCIR-14 Short Text Conversation Task (STC-3)

- [NTCIR](#)
- [Twitter: @ntcirstc](#)
- [STC-3@NTCIR-14](#)

Welcome to the top page of STC-3@NTCIR-14!
STC-3 offers three subtasks:

- [Chinese Emotional Conversation Generation \(CECG\) Subtask](#)
- Dialogue Quality (DQ) Subtask (for Chinese and English)
- Nugget Detection (ND) Subtask (for Chinese and English)

Key dates for DQ and ND Subtasks

Feb-Mar 2018 Crawling Chinese test data from Weibo

Oct 2017-Jan 2018 Training data translation into English

Apr-Jun, 2018 Test data translation into English

Jul-Aug 2018 Training/test data annotation

Aug 31, 2018 STC-3 task registrations due (CECG, DQ, ND)

Sep 1, 2018 Training data with annotations released

Nov 1, 2018 Test data released

Nov 30, 2018 Run submissions due

Dec 20, 2018 Results and draft overview released to participants

Feb 1, 2019 Participant papers due

Mar 1, 2019 Acceptance notification

Mar 20, 2019 All camera-ready papers due

Jun 2019 NTCIR-14 Conference@NII

NTCIR-14 STC-3

Short Text Conversation Task (STC-3)

Chinese Emotional Conversation Generation (CECG) Subtask



Short Text Conversation Task (STC-3)

Chinese Emotional Conversation Generation (CECG) Subtask

Home

Task Definition

Dataset Description

Evaluation Metric

Time Schedule

Copy Rights &
Contacts

Call for Participation

In recent years, there has been a rising tendency in AI research to enhance Human-Computer Interaction by humanizing machines. However, to create a robot capable of acting and talking with a user at the human level requires the robot to understand human cognitive behaviors, while one of the most important human behaviors is expressing and understanding emotions and affects. As a vital part of human intelligence, emotional intelligence is defined as the ability to perceive, integrate, understand, and regulate emotions. Though a variety of models have been proposed for conversation generation from large-scale social data, it is still quite challenging (and yet to be addressed) to generate emotional responses.

In this challenge, participants are expected to generate Chinese responses that are not only appropriate in content but also adequate in emotion, which is quite important for building an empathic chatting machine. For instance, if user says “My cat died yesterday”, the most appropriate response may be “It’s so sad, so sorry to hear that” to express sadness, but also could be “Bad things always happen, I hope you will be happy soon” to express comfort.

[Previous Evaluation Challenge at NLPCC 2017](#)

[Overview of the NLPCC 2017 Shared Task: Emotion Generation Challenge](#)

Links

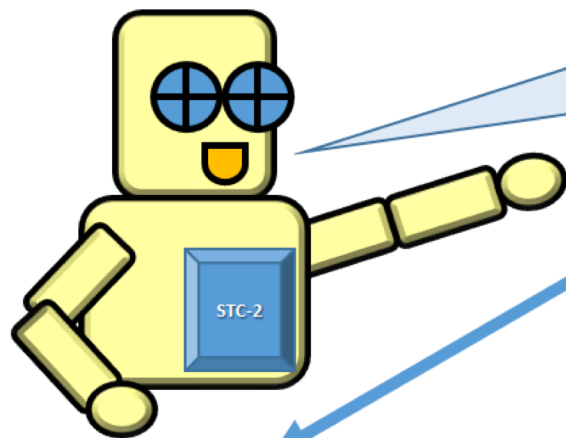
[NTCIR](#) NTCIR-14

[STC-3](#) NTCIR-14 STC-3

[NLPCC](#) NLPCC 2017

Short Text Conversation (NTCIR-13 STC2) Retrieval-based

retrieval-based method



Given a new post, can a **coherent** and **useful** comment be returned by searching a post-comment repository?

post

Search and reuse

post-comment repository

post

comment

comment

post

comment

comment

post

comment

comment

post

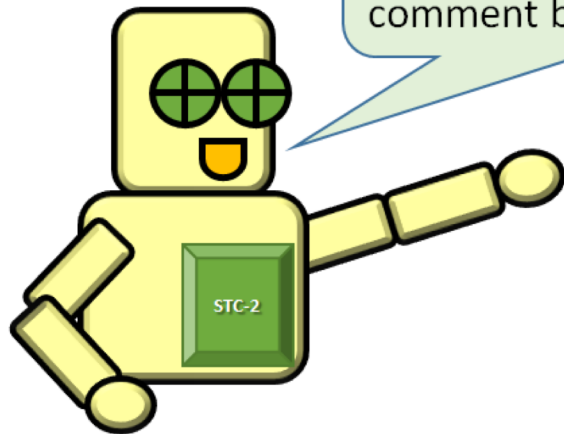
comment

comment

Short Text Conversation (NTCIR-13 STC2) Generation-based

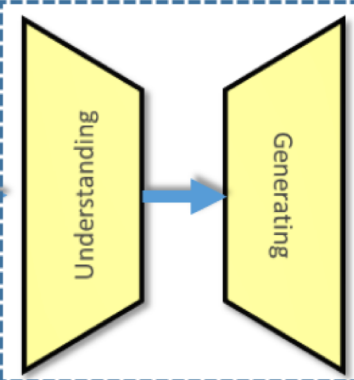
generation-based method

Given a new post, can a
fluent, coherent and useful
comment be generated?



post

The Trained Generator



generated comment
generated comment
generated comment

Used to train the generator

post-comment repository

post

comment

comment

post

comment

comment

post

comment

comment

post

comment

comment

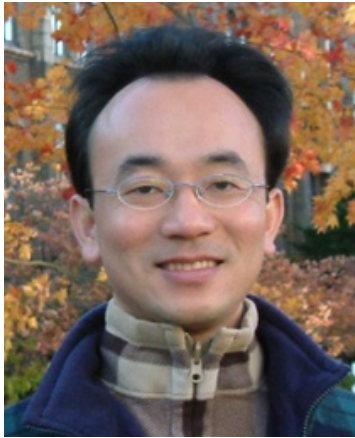
Summary



- This course introduces the **fundamental concepts** and **research issues** of **Big Data Mining**.
- Topics include
 - ABC: AI, Big Data, Cloud Computing,
 - Data Science and Big Data Analytics: Discovering, Analyzing, Visualizing and Presenting Data,
 - Fundamental Big Data: MapReduce Paradigm, Hadoop and Spark Ecosystem,
 - Foundations of Big Data Mining in Python,
 - Supervised Learning: Classification and Prediction,
 - Unsupervised Learning: Cluster Analysis,
 - Unsupervised Learning: Association Analysis,
 - Machine Learning with Scikit-Learn in Python,
 - Deep Learning for Finance Big Data with TensorFlow,
 - Convolutional Neural Networks (CNN)
 - Recurrent Neural Networks (RNN)
 - Reinforcement Learning (RL)
 - Social Network Analysis (SNA)



Big Data Mining Contact



Min-Yuh Day, Ph.D.

Assistant Professor

[Department of Information Management,
Tamkang University](#)

Tel: 886-2-26215656 ext. 2846

Fax: 886-2-26209737

Office: B929

Address: No.151, Yingzhuang Rd., Danshui Dist.,
New Taipei City 25137, Taiwan (R.O.C.)

Email: myday@mail.tku.edu.tw

Web: <http://mail.tku.edu.tw/myday/>

