Tamkang University

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

Tamkang University

- 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)

Tamkang University

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)

Tamkang University

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

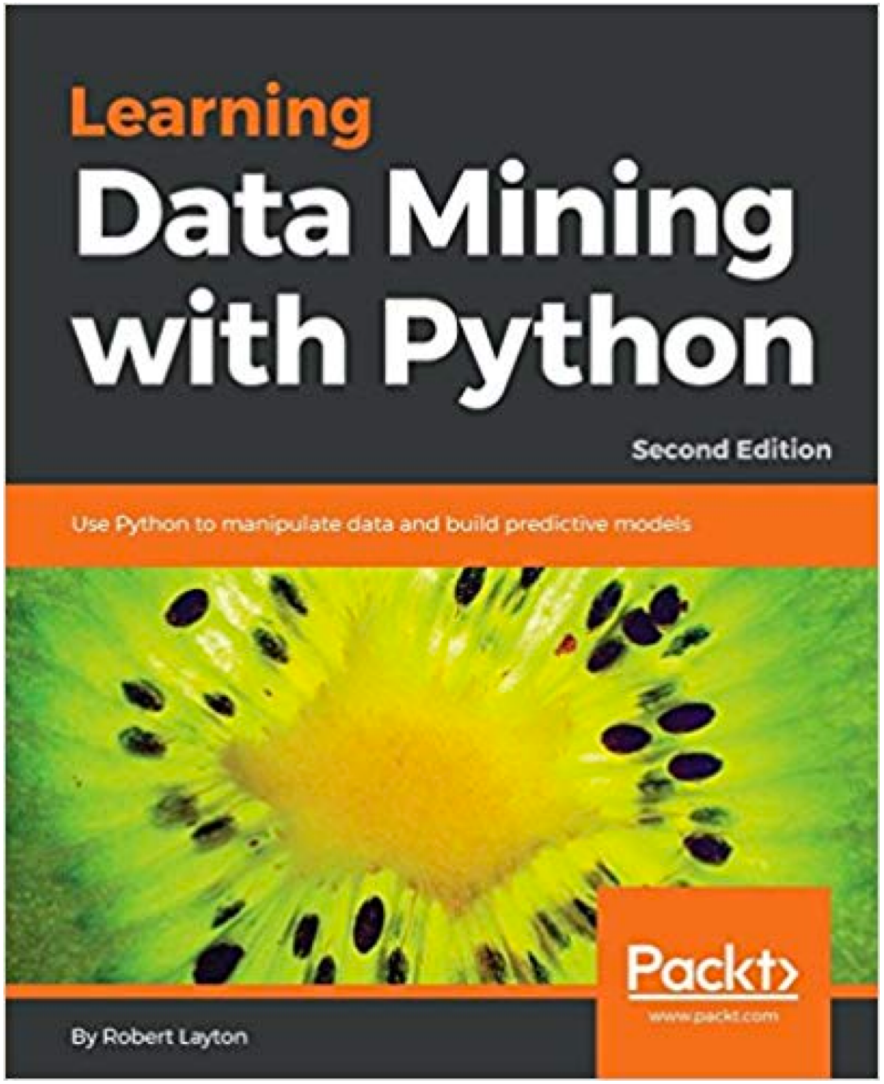
Grading Policy

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

Textbooks and References

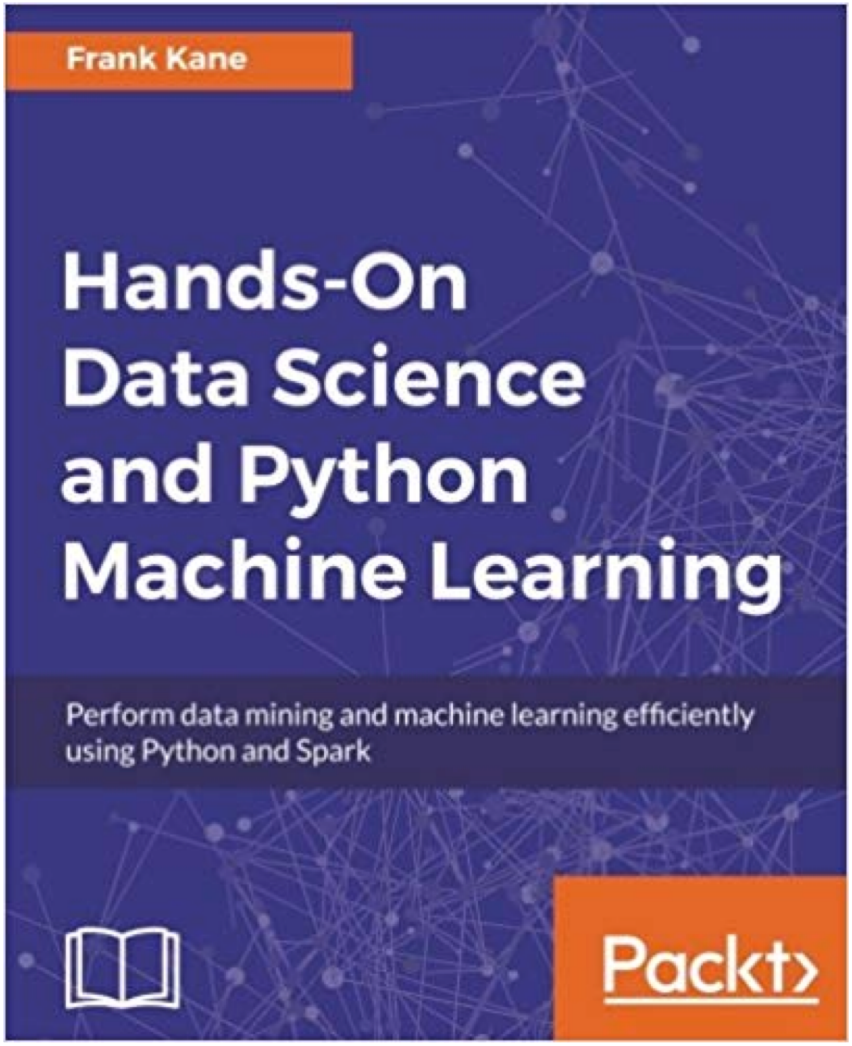
- Textbook: Slides
  - http://mail.tku.edu.tw/myday/teaching.htm#1071BDM
- References
  - Learning Data Mining with Python - Second Edition, Robert Layton, Packt Publishing, 2017.
  - Hands-On Data Science and Python Machine Learning: Perform data mining and machine learning efficiently using Python 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, MichelineKamber and Jian Pei, Morgan Kaufmann, 2011.

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



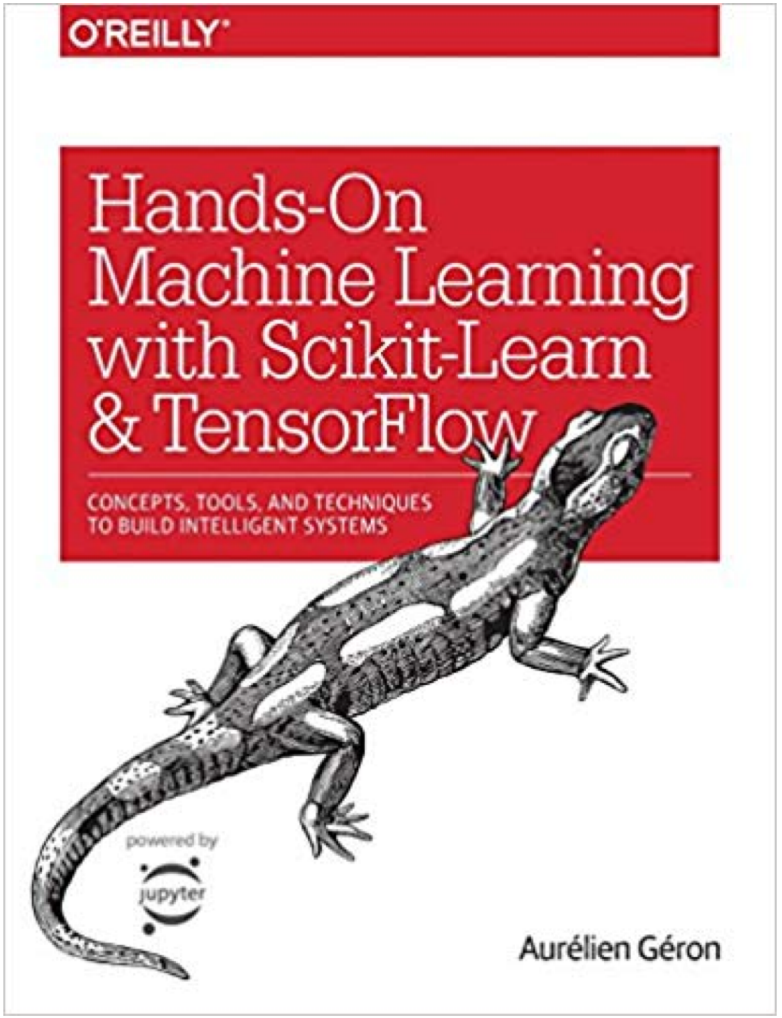
Source: https://www.amazon.com/Learning-Data-Mining-Python-Second/dp/1787126781

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



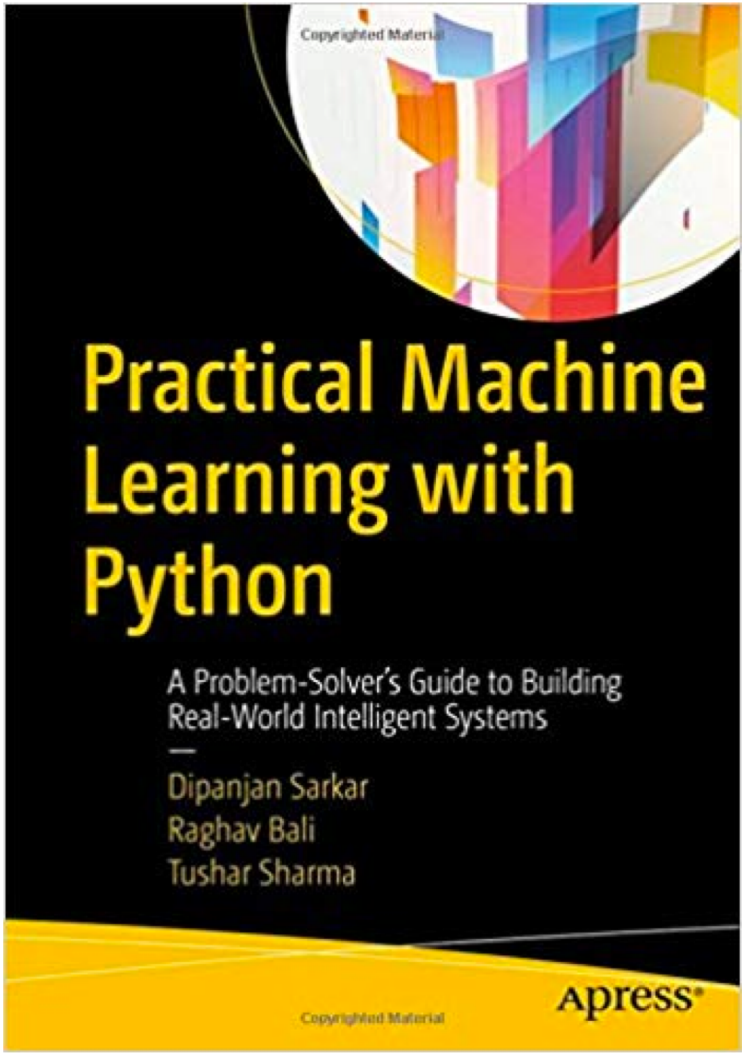
Source: https://www.amazon.com/Hands-Science-Python-Machine-Learning/dp/1787280748

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



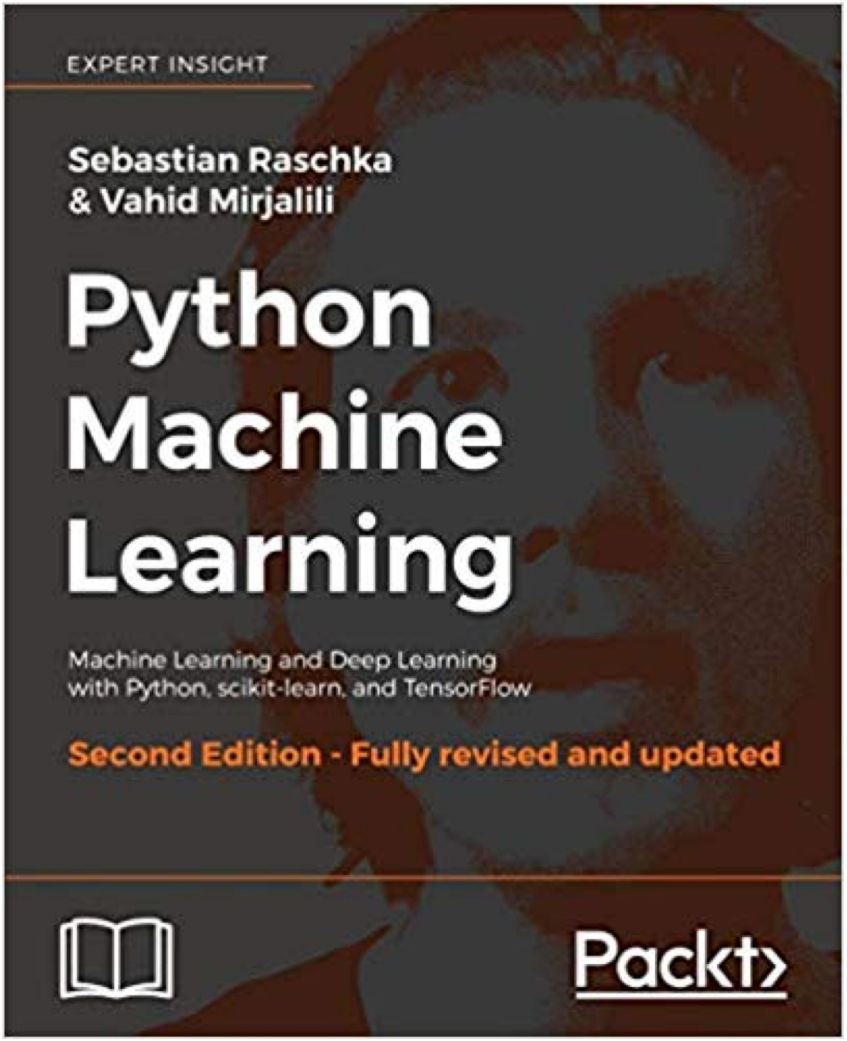
Source: https://www.amazon.com/Hands-Machine-Learning-Scikit-Learn-TensorFlow/dp/1491962291

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



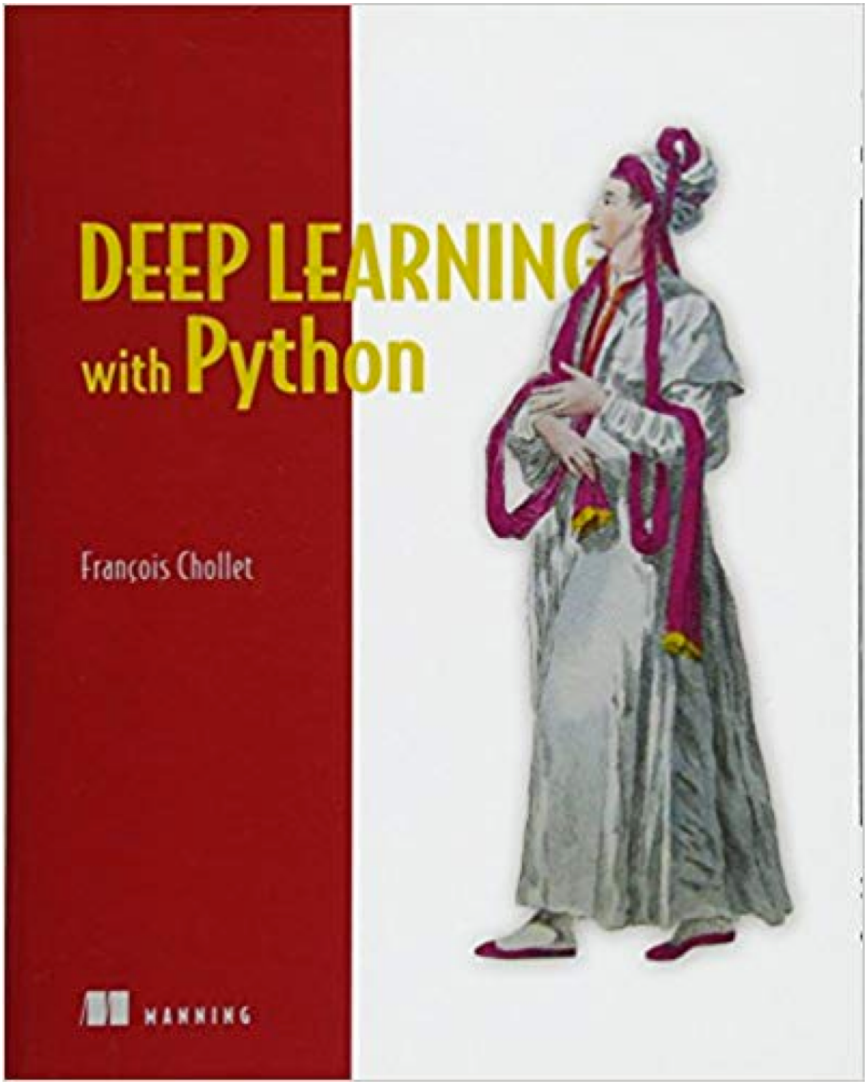
Source: https://www.amazon.com/Practical-Machine-Learning-Python-Problem-Solvers/dp/1484232062

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



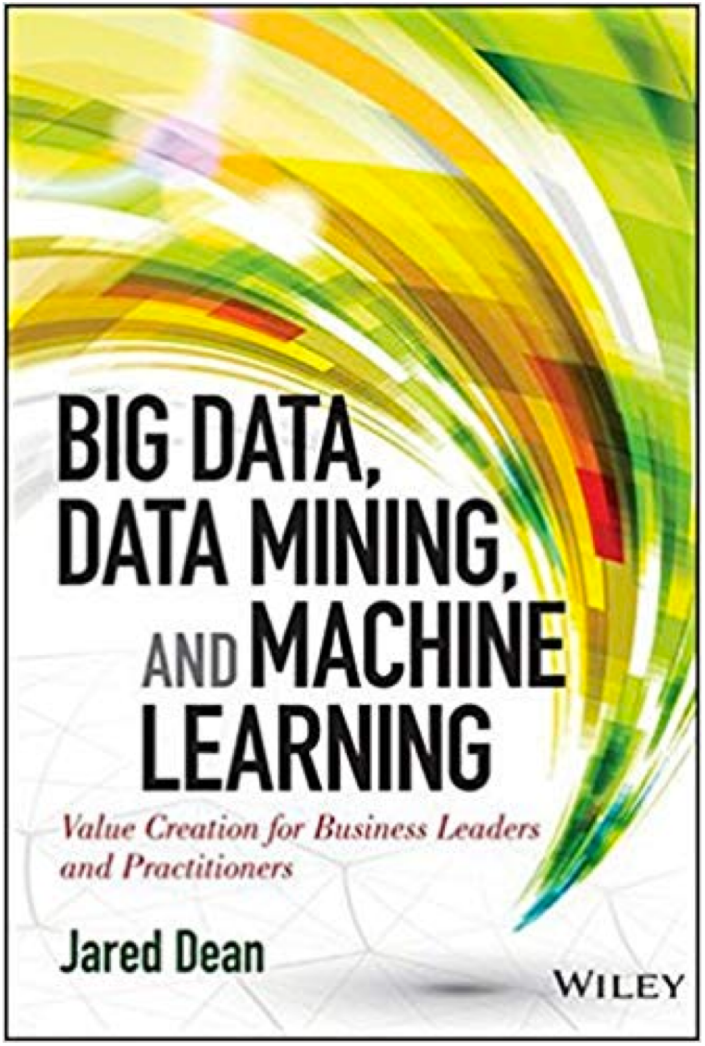
Source: https://www.amazon.com/Python-Machine-Learning-scikit-learn-TensorFlow/dp/1787125939

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



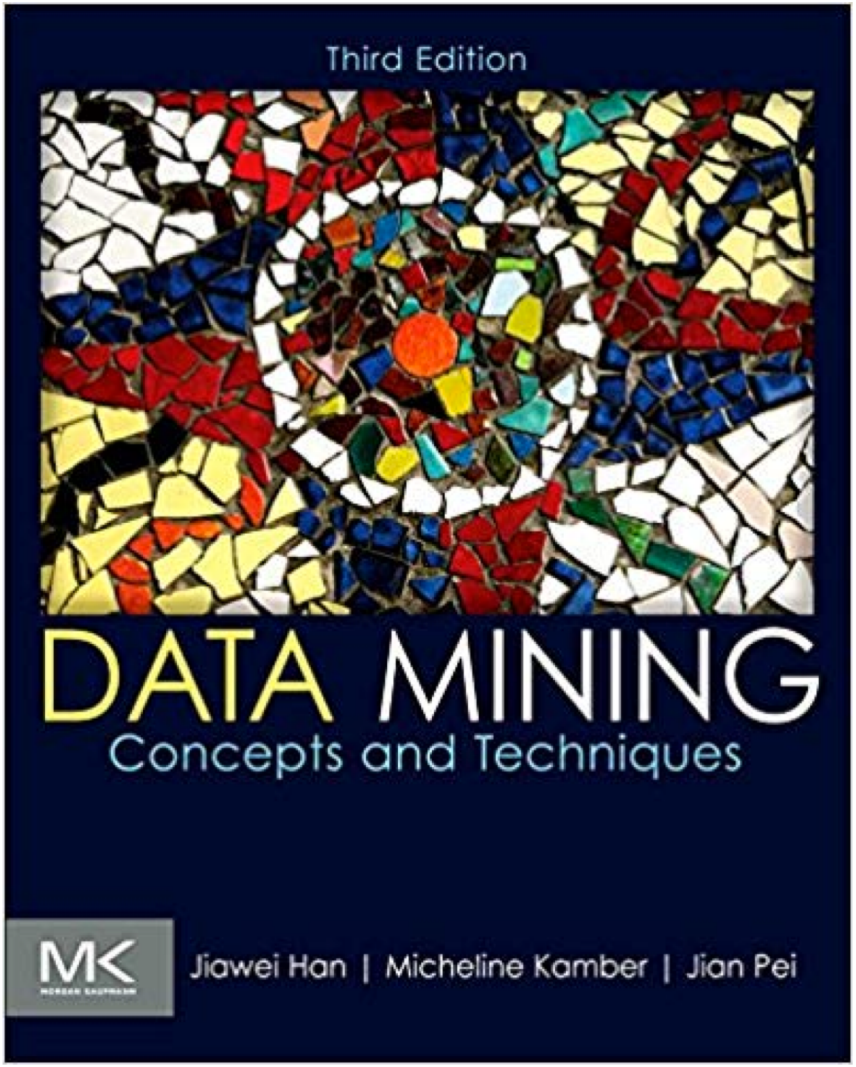
Source: https://www.amazon.com/Deep-Learning-Python-Francois-Chollet/dp/1617294438

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



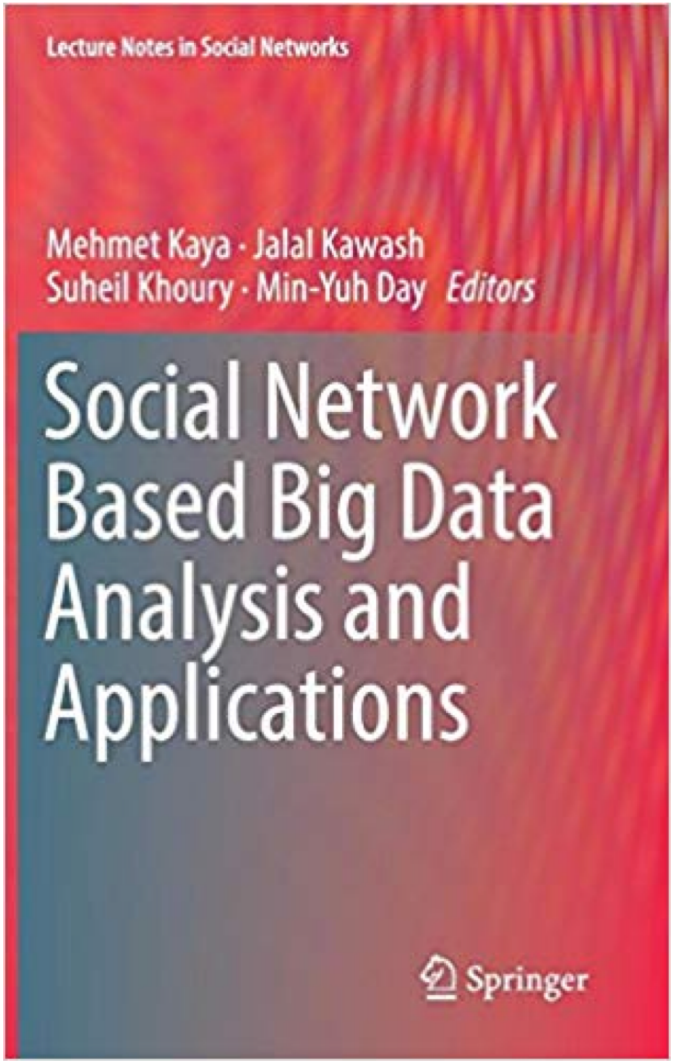
Source: https://www.amazon.com/Data-Mining-Machine-Learning-Practitioners/dp/1118618041

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



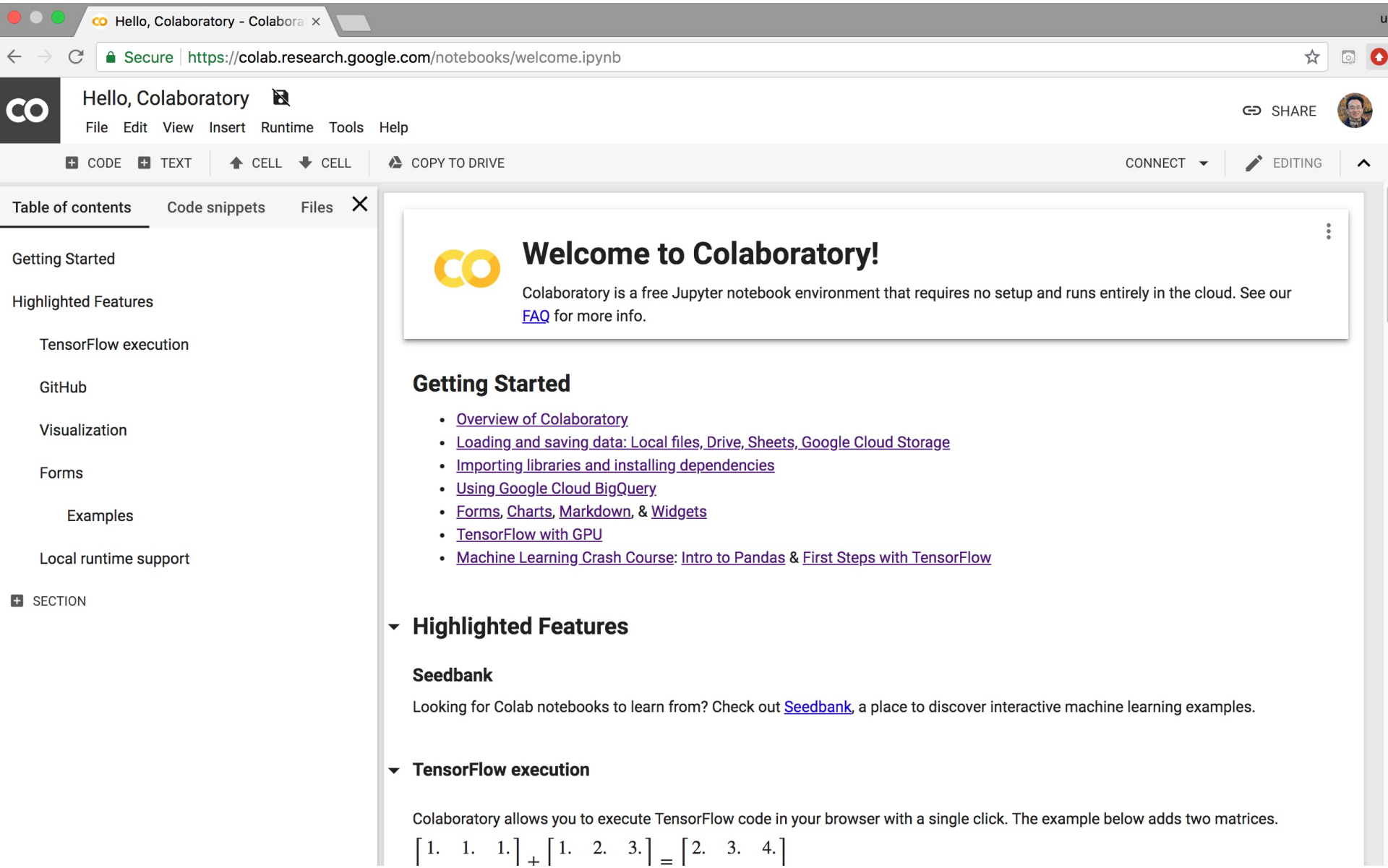
Source: https://www.amazon.com/Data-Mining-Concepts-Techniques-Management/dp/0123814790

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.



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

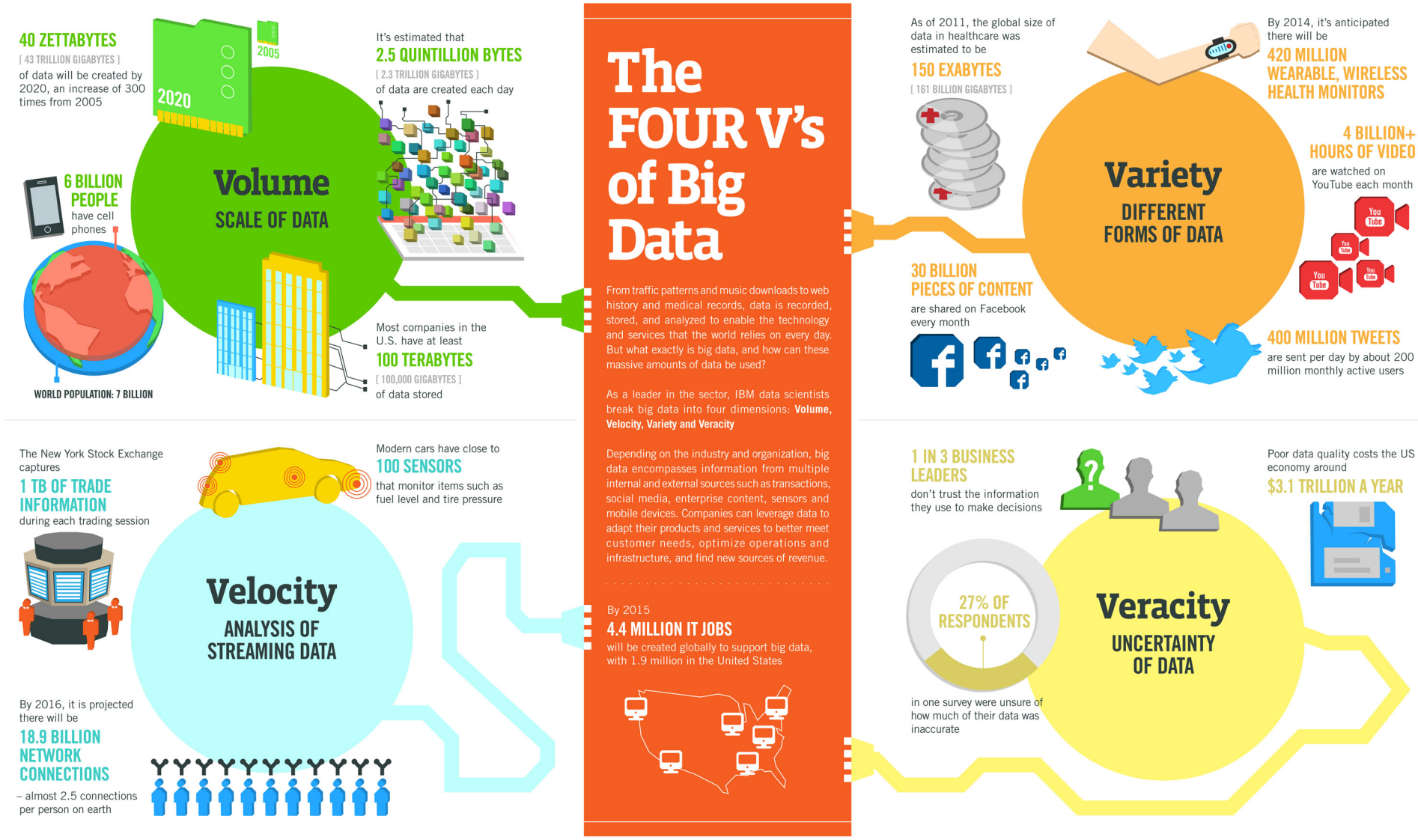
Google Colab



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

Big Data Analytics and Data Mining

Big Data 4 V



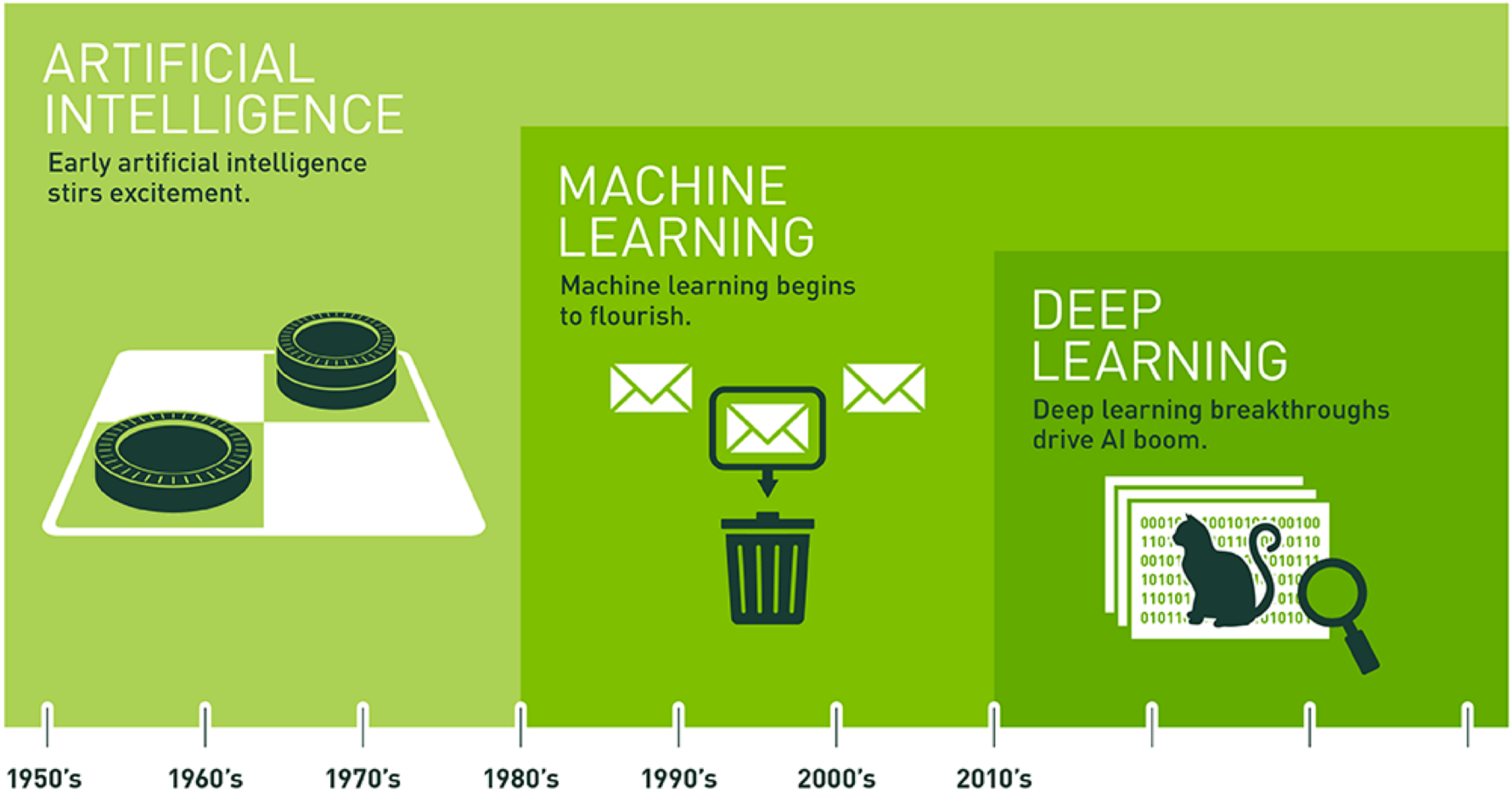
Sources: McKinsey Global Institute, Twitter, Cisco, Gartner, EMC, SAS, IBM, MEPTEC, QAS

IBM

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

Value

Artificial Intelligence
Machine Learning & Deep Learning

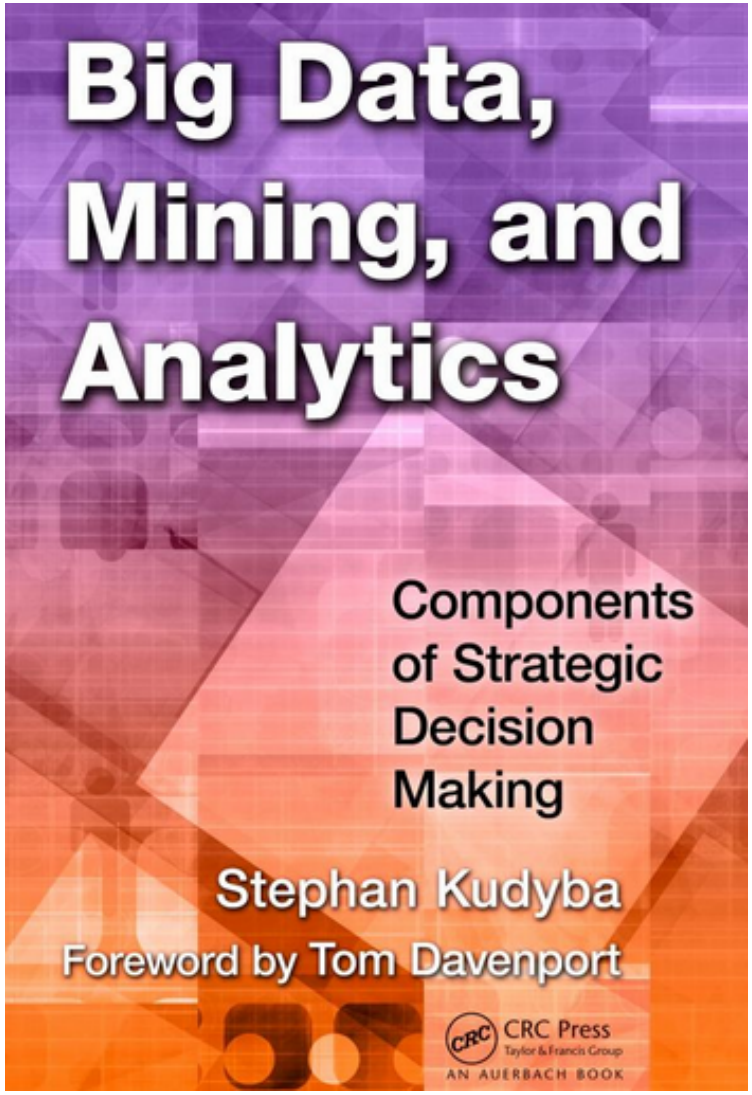


Since an early flush of optimism in the 1950s, smaller subsets of artificial intelligence – first machine learning, then deep learning, a subset of machine learning – have created ever larger disruptions.

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

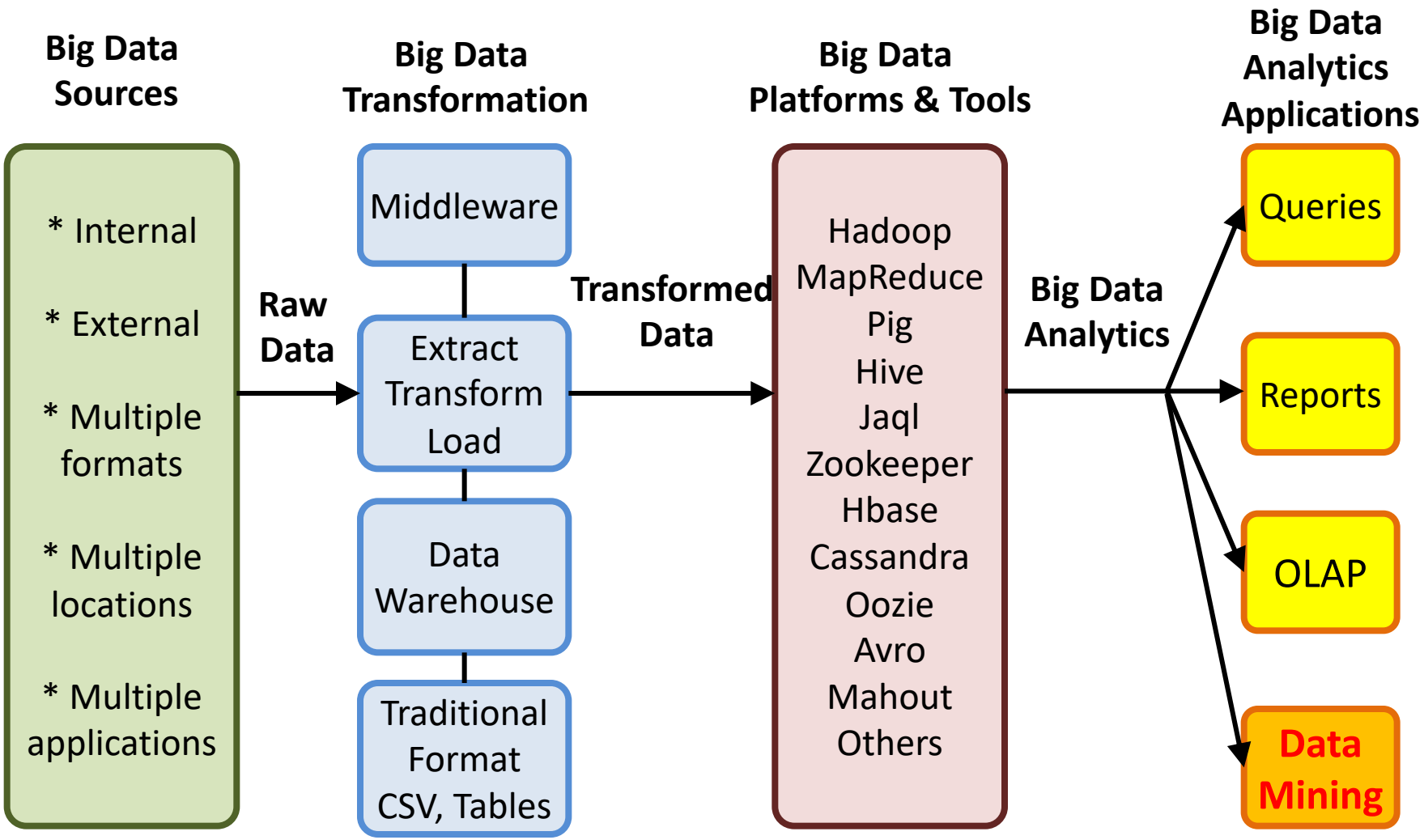
Stephan Kudyba (2014),

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



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

Architecture of Big Data Analytics



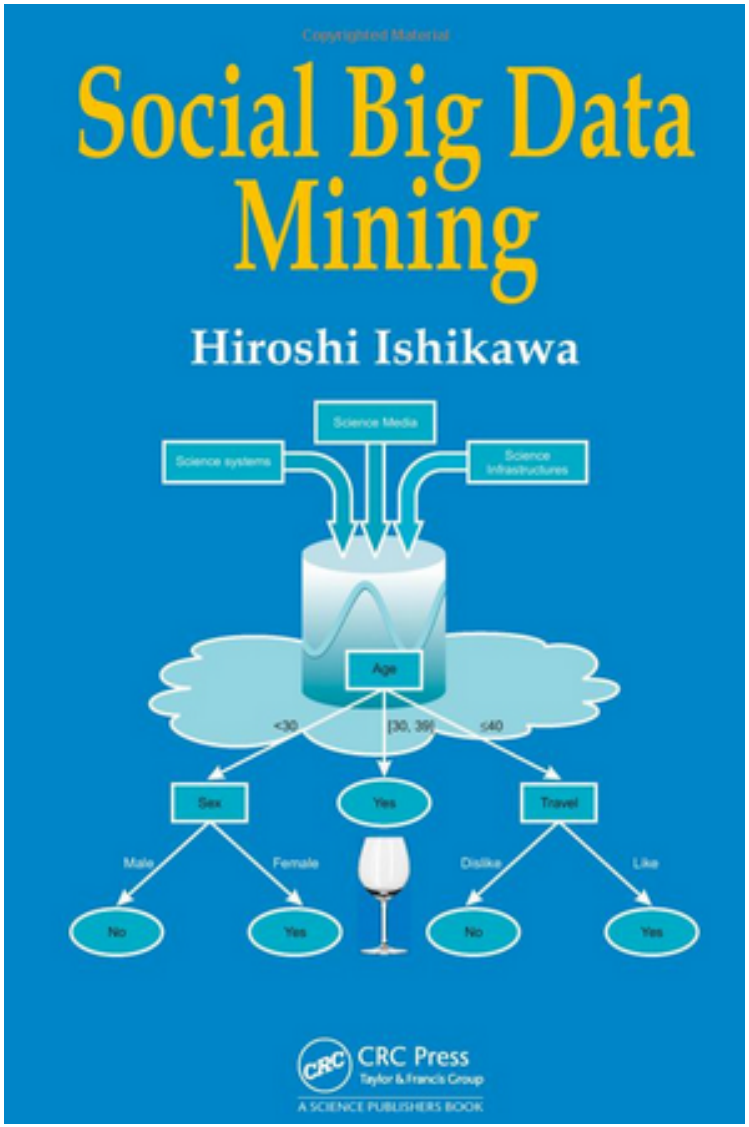
Source: Stephan Kudyba (2014), Big Data, Mining, and Analytics: Components of Strategic Decision Making, Auerbach Publications

Architecture of Big Data Analytics

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

Social Big Data Mining

(Hiroshi Ishikawa, 2015)



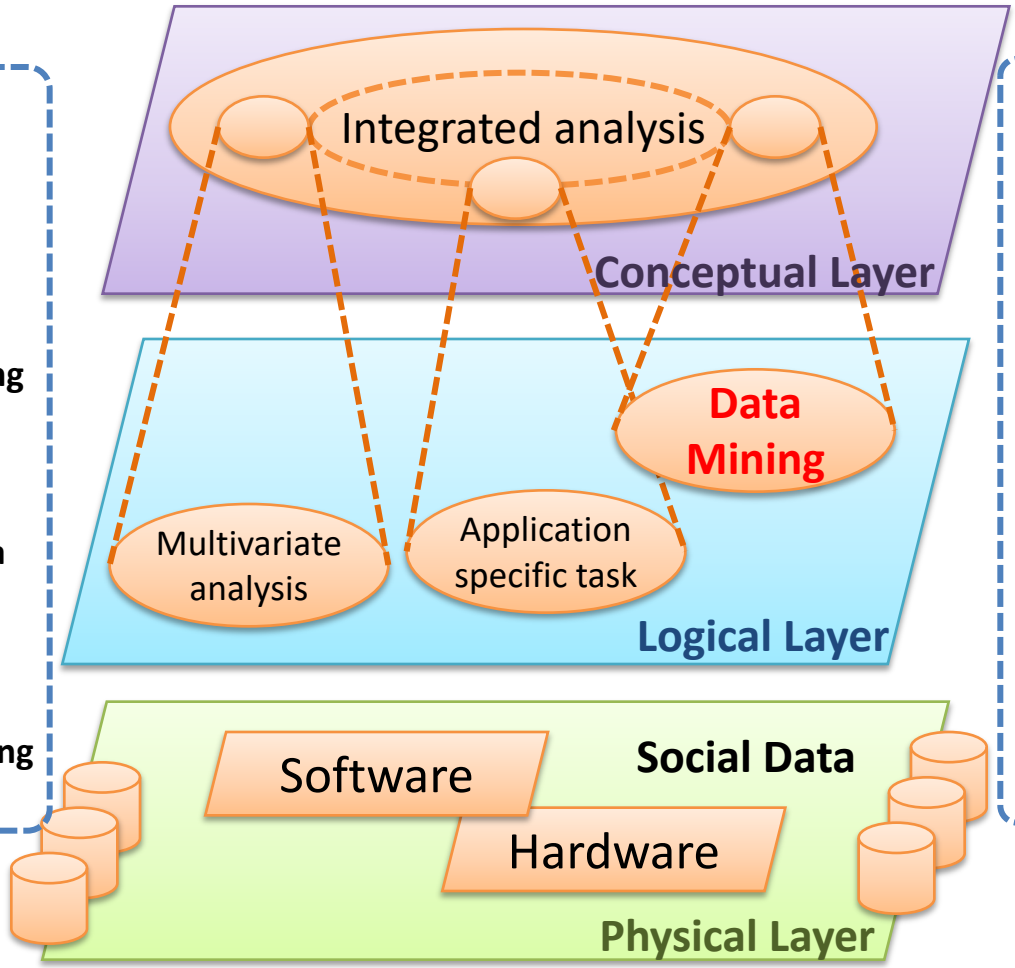
Source: http://www.amazon.com/Social-Data-Mining-Hiroshi-Ishikawa/dp/149871093X

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

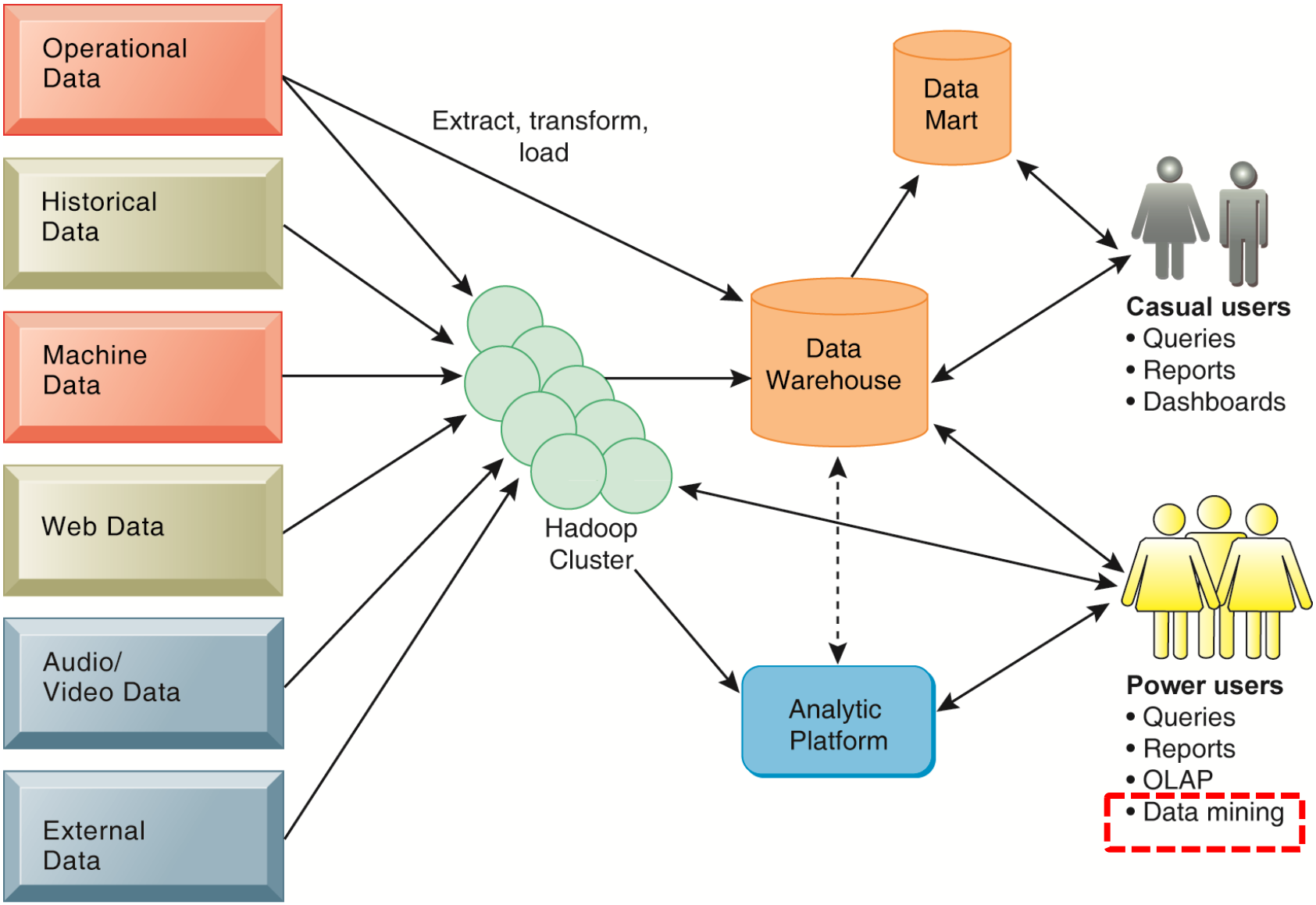


Analysts

- 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

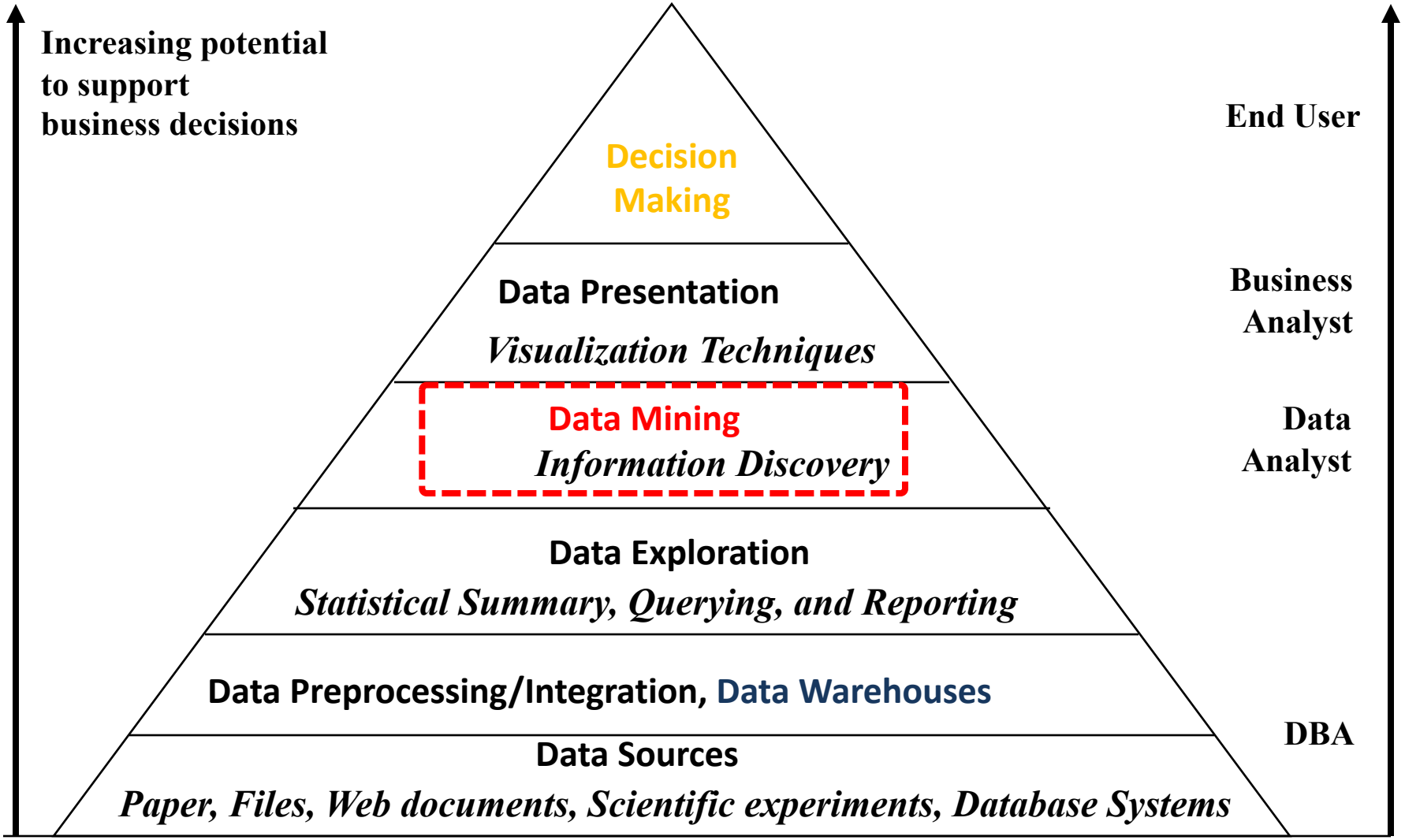
Business Intelligence (BI) Infrastructure



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

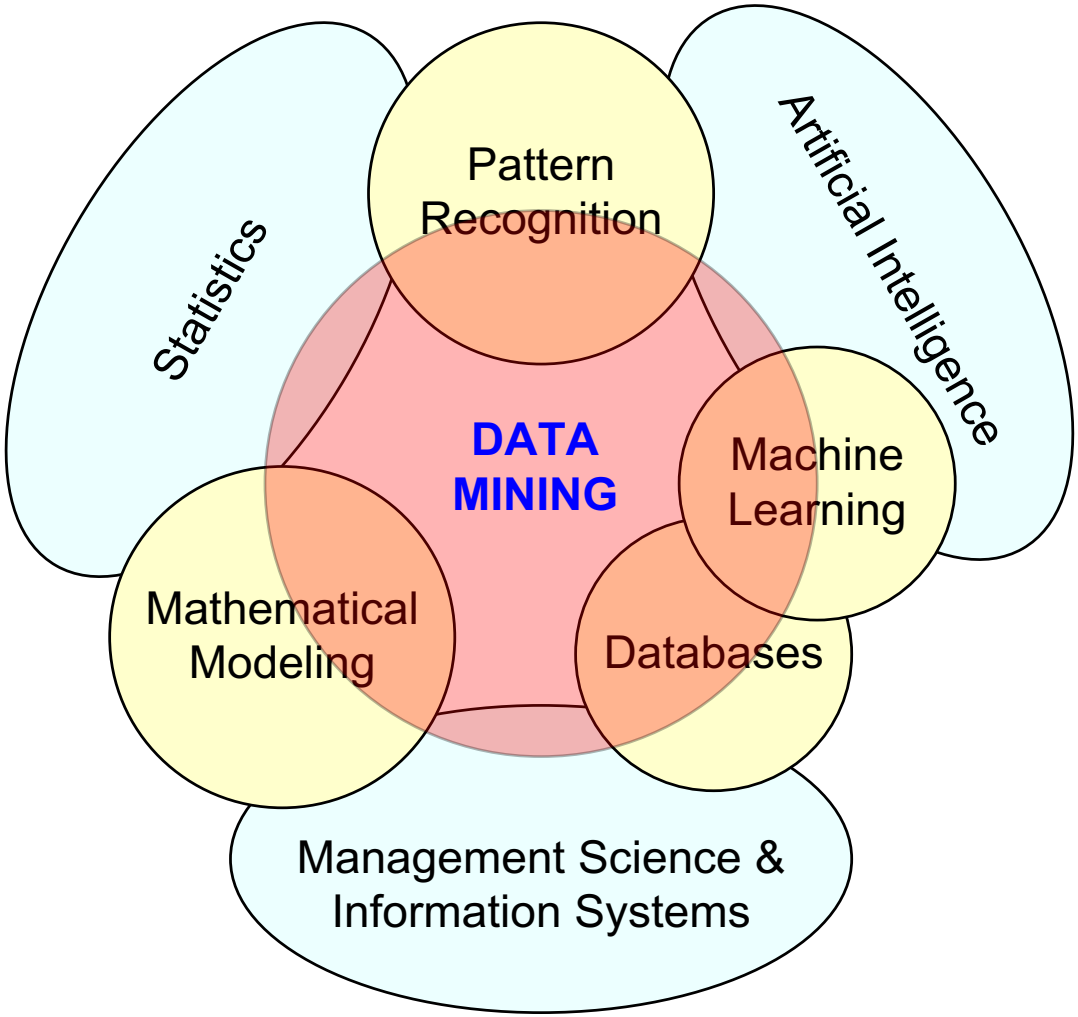


Source: Jiawei Han and Micheline Kamber (2006), Data Mining: Concepts and Techniques, Second Edition, Elsevier

The Evolution of BI Capabilities

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

Data Mining at the Intersection of Many Disciplines



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

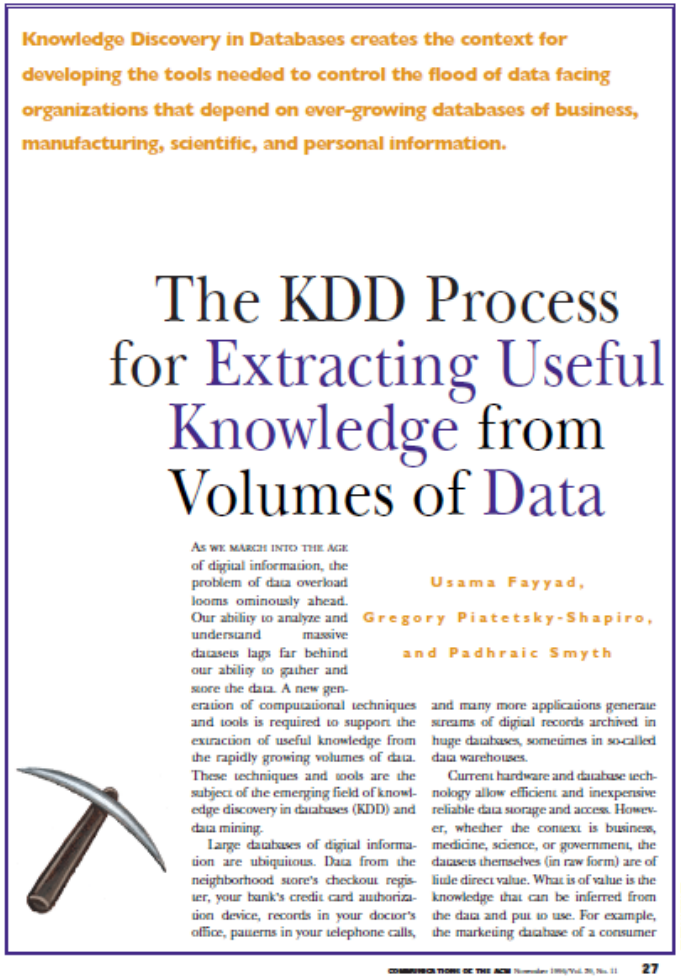
Data Mining:

Core Analytics Process

The KDD Process for Extracting Useful Knowledge from Volumes of Data

Source: 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.

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

Usama Fayyad,
Gregory Piatetsky-Shapiro,
and Padhraic Smyth

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.

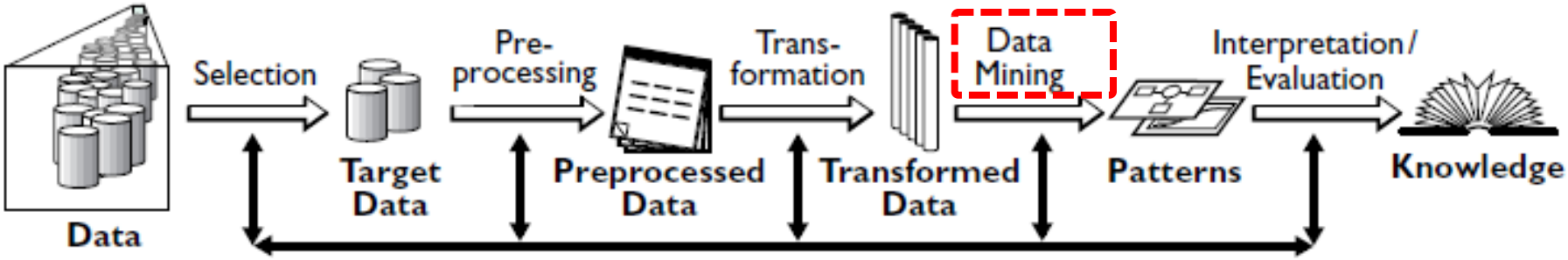
Current hardware and database technology allow efficient and inexpensive reliable data storage and access. However, whether the context is business, medicine, science, or government, the datasets themselves (in raw form) are of little direct value. What is of value is the knowledge that can be inferred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

TERRY WEEHER

Data Mining

Knowledge Discovery in Databases (KDD) Process

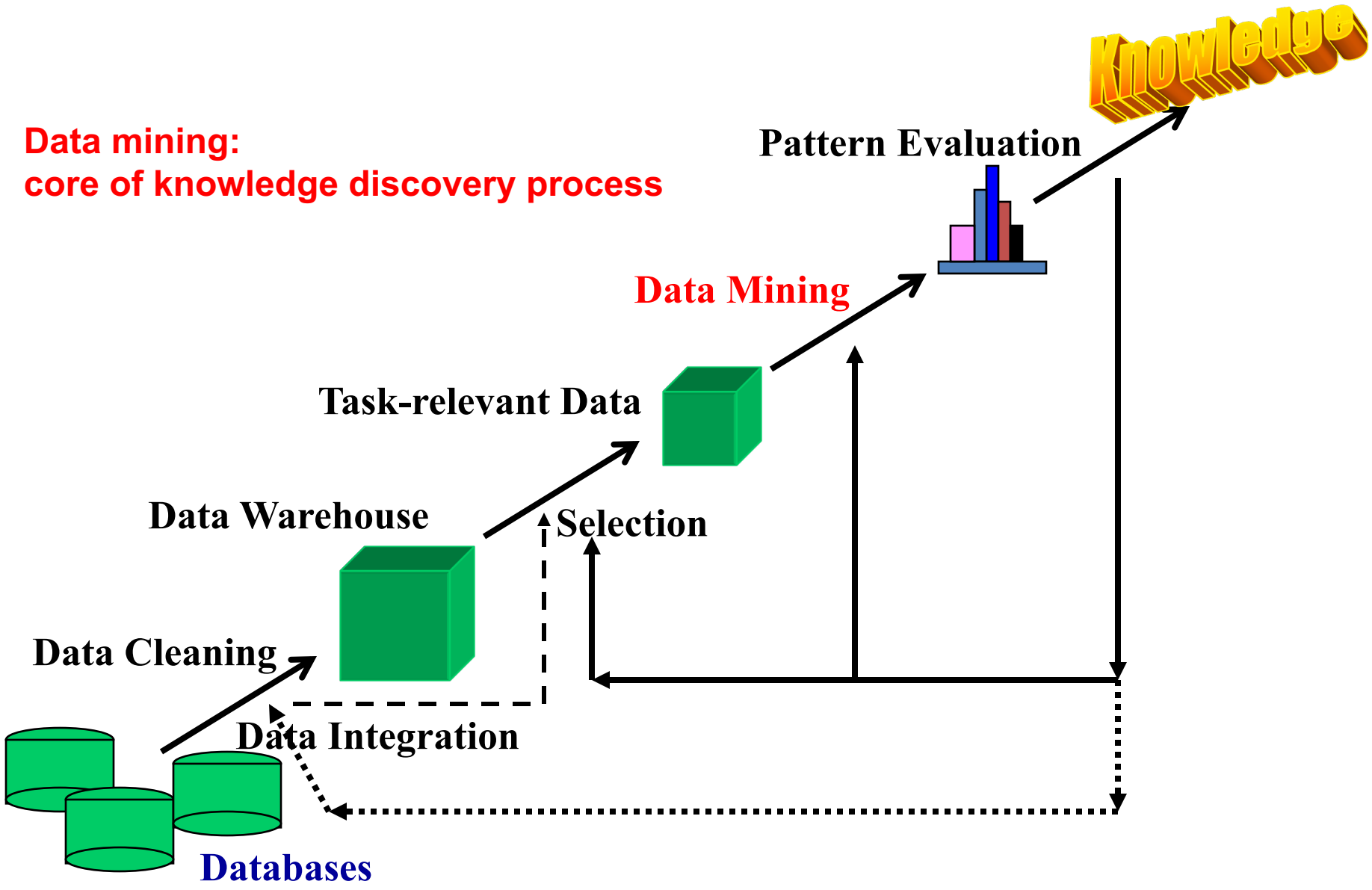
(Fayyad et al., 1996)



Source: 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 (KDD) Process

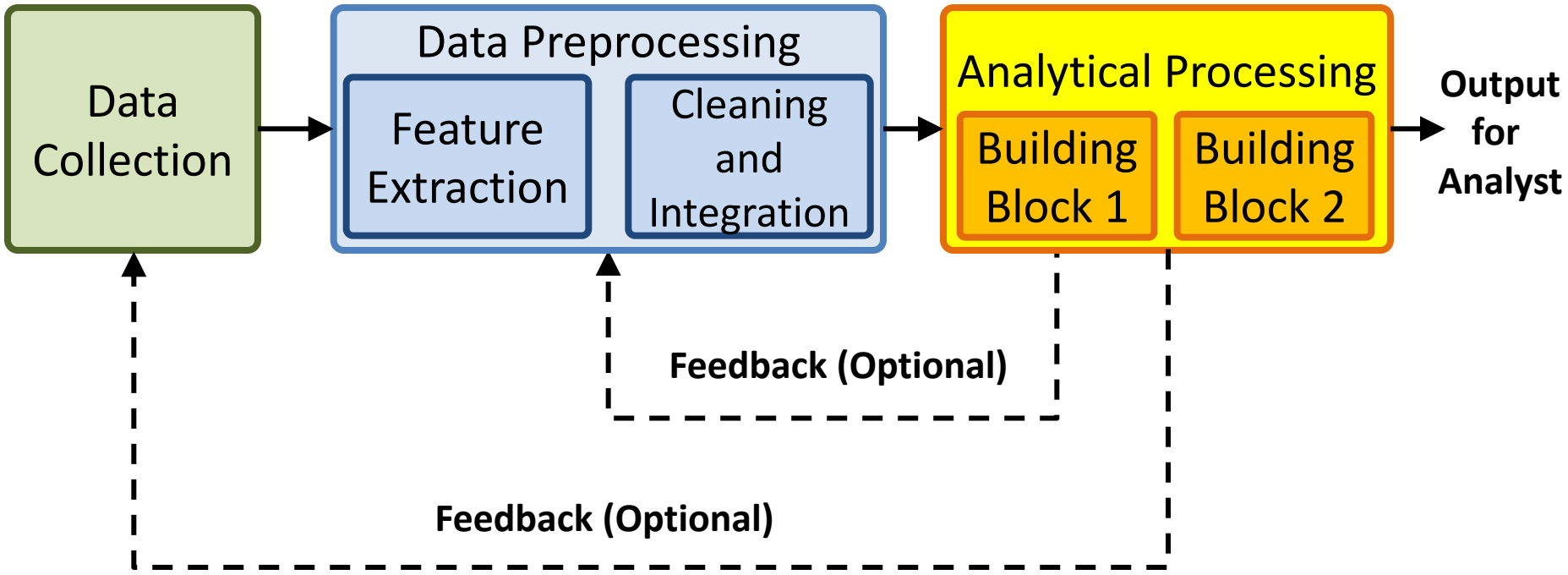
Data mining:
core of knowledge discovery process



Source: Han & Kamber (2006)

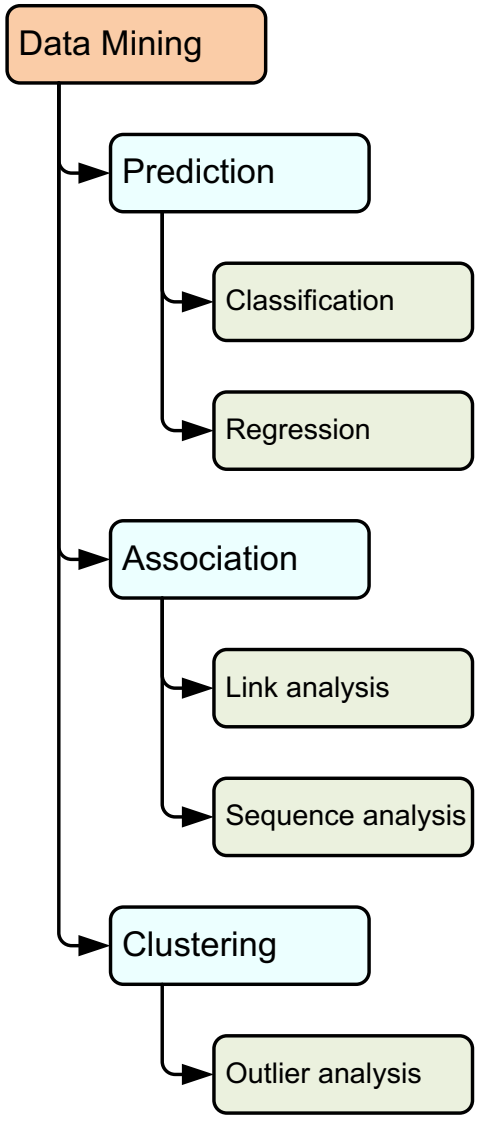
Data Mining Processing Pipeline

(Charu Aggarwal, 2015)



Source: Charu Aggarwal (2015), Data Mining: The Textbook Hardcover, Springer

A Taxonomy for Data Mining Tasks



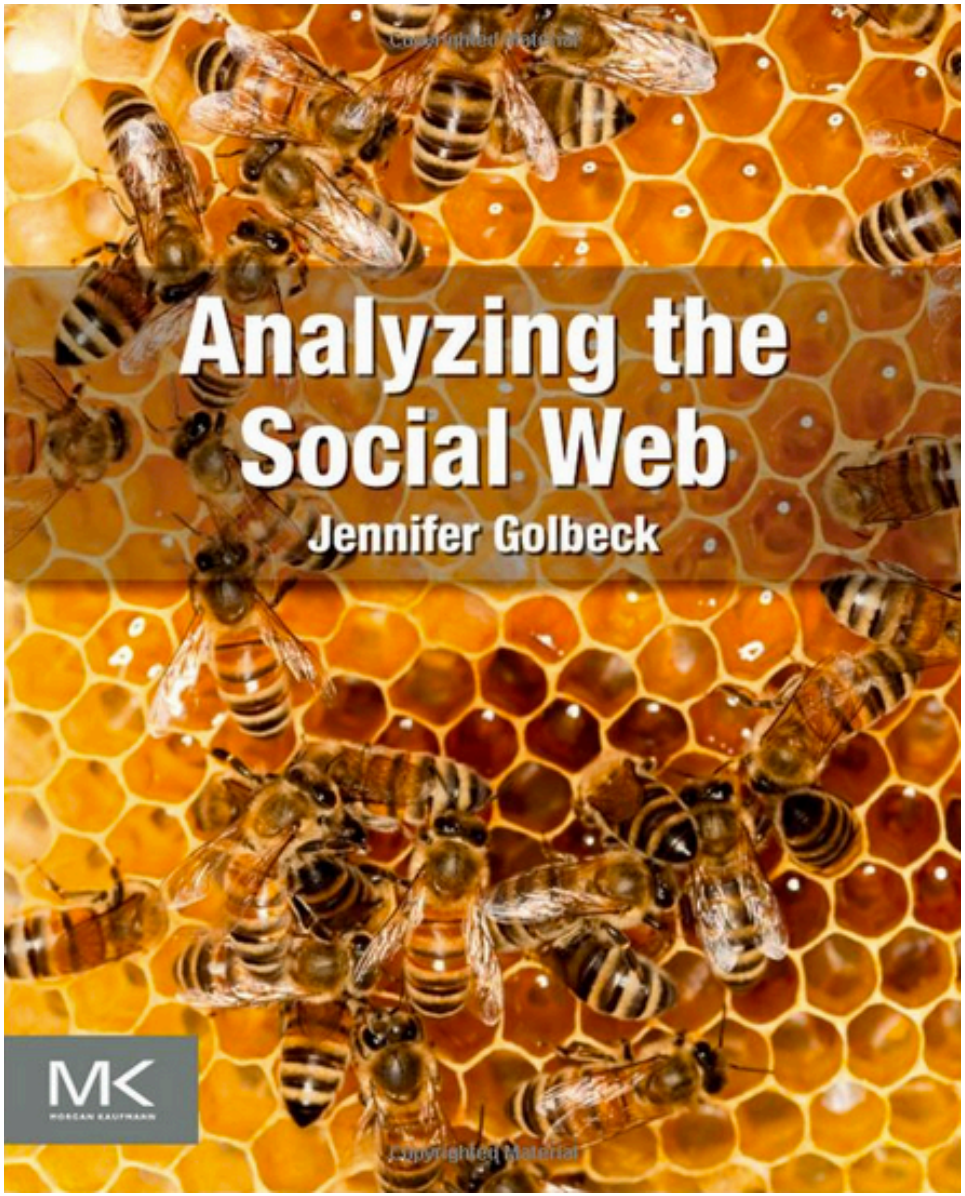
Learning Method	Popular Algorithms
Supervised	Classification and Regression Trees, ANN, SVM, Genetic Algorithms
Supervised	Decision trees, ANN/MLP, SVM, Rough sets, Genetic Algorithms
Supervised	Linear/Nonlinear Regression, Regression trees, ANN/MLP, SVM
Unsupervised	Apriory, OneR, ZeroR, Eclat
Unsupervised	Expectation Maximization, Apriory Algorithm, Graph-based Matching
Unsupervised	Apriory Algorithm, FP-Growth technique
Unsupervised	K-means, ANN/SOM
Unsupervised	K-means, Expectation Maximization (EM)

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

Business Insights with Social Analytics

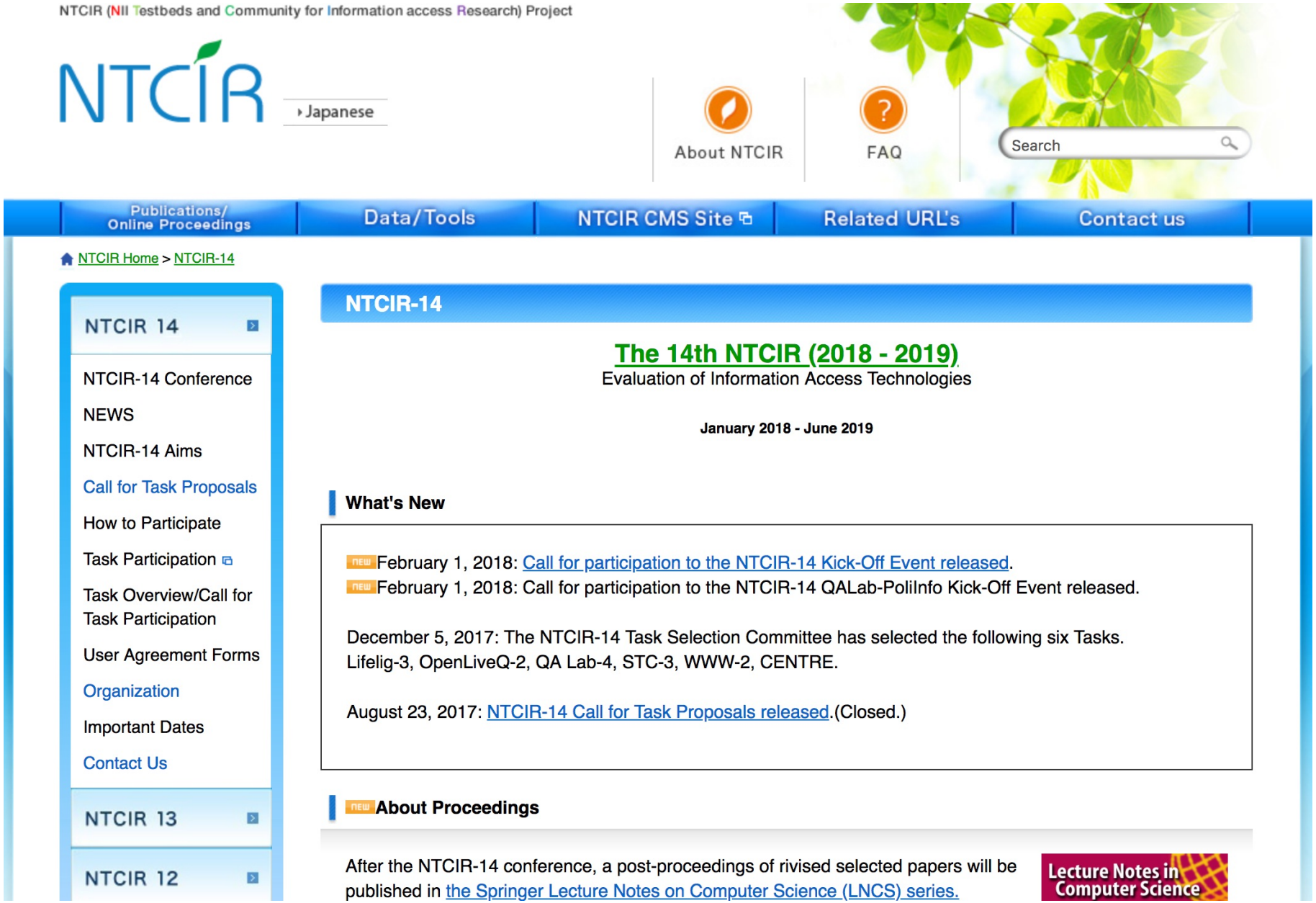
Analyzing the Social Web:
Social Network Analysis

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



Source: http://www.amazon.com/Analyzing-Social-Web-Jennifer-Golbeck/dp/0124055311

The 14th NTCIR (2018 - 2019)



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

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 2018Crawling 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

NTCIR

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

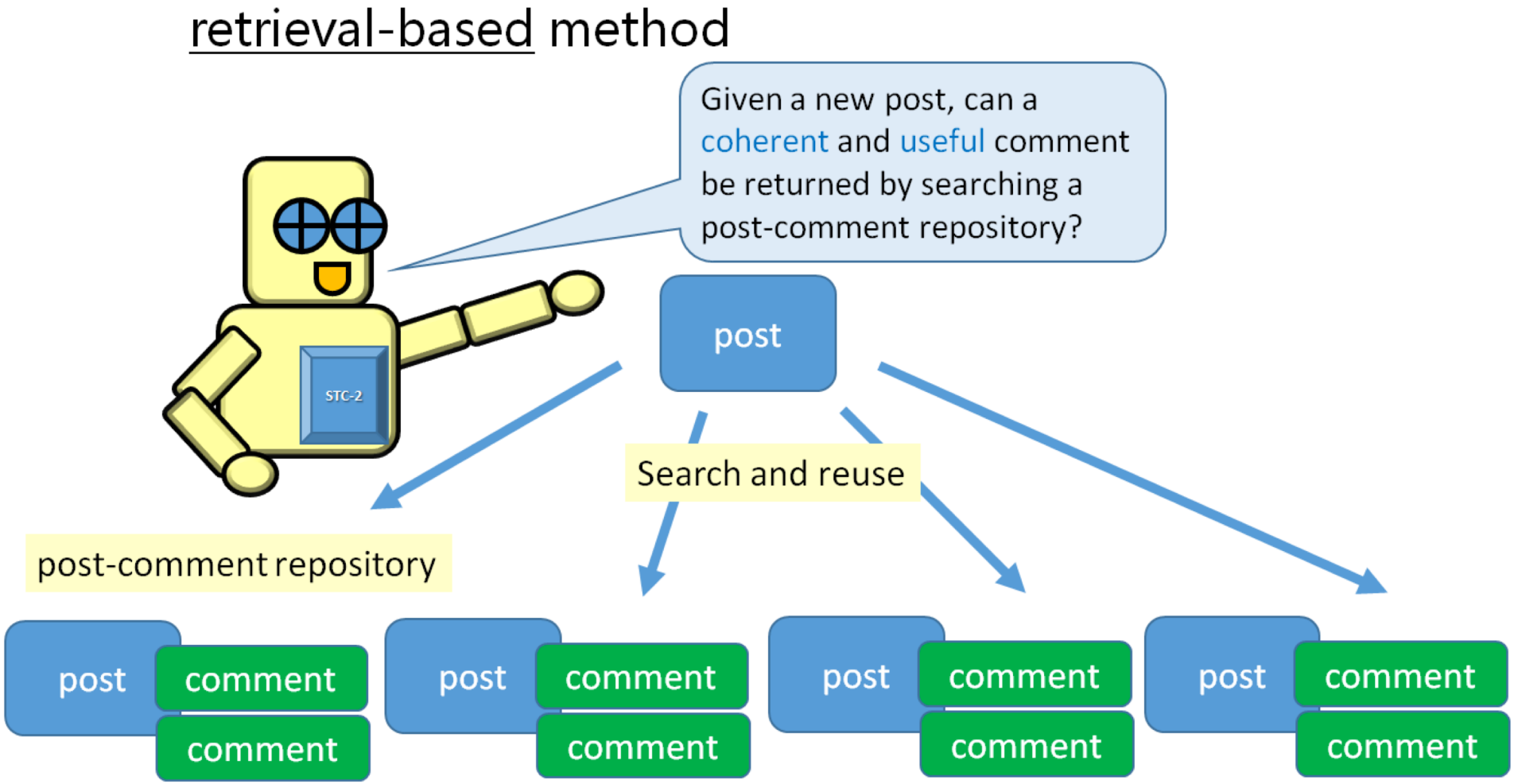
STC-3 NTCIR-14 STC-3

NLPCC NLPCC 2017

AI lab. of Computer Science, Tsinghua University, Beijing 100084, China.

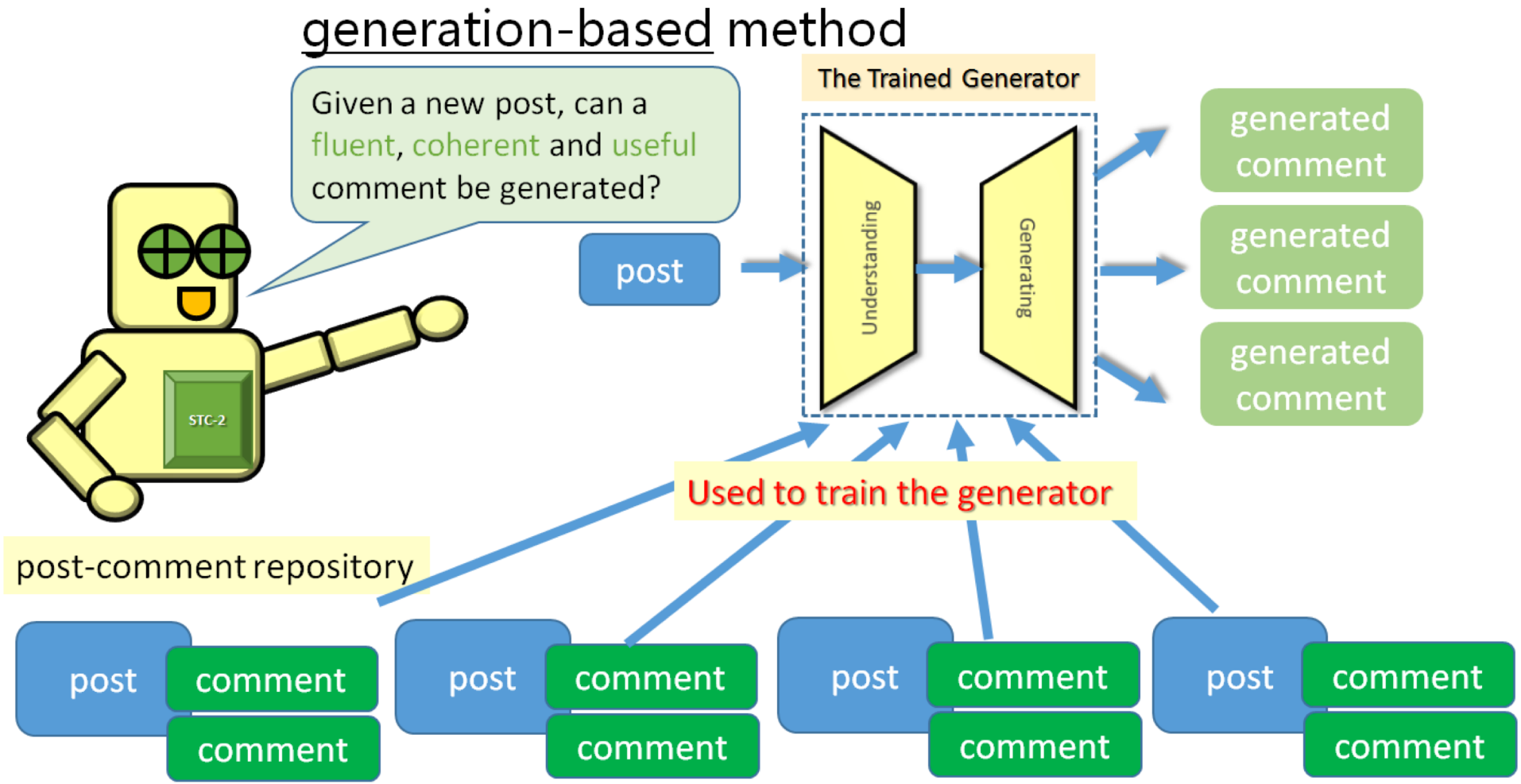
http://www.aihuang.org/p/challenge.html

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



Source: http://ntcirstc.noahlab.com.hk/STC2/stc-cn.htm

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



Source: http://ntcirstc.noahlab.com.hk/STC2/stc-cn.htm

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)

Tamkang University

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, Yingzhuan Rd., Danshui Dist., New Taipei City 25137, Taiwan (R.O.C.)

Email: myday@mail.tku.edu.tw

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