Tamkang University
淡江大學

Tamkang University

Big Data Mining
巨量資料探勘
Course Orientation for Big Data Mining
(巨量資料探勘課程介紹)

1062DM01
MI4 (M2244) (2995)
Wed, 9, 10 (16:10-18:00) (B206)

Min-Yuh Day
戴敏育
Assistant Professor
專任助理教授

Dept. of Information Management, Tamkang University
淡江大學 資訊管理學系

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

2018-03-07

淡江大學106學年度第2學期
課程教學計畫表
Spring 2018 (2018.02 - 2018.06)

- 課程名稱：巨量資料探勘 (Big Data Mining)
- 授課教師：戴敏育 (Min-Yuh Day)
- 開課系級：資管四P (TLMXB4P) (M2244) (2995)
- 開課資料：選修 單學期 2 學分 (2 Credits, Elective)
- 上課時間：週三 9,10 (Wed 16:10-18:00)
- 上課教室： B206

課程簡介

- 本課程介紹巨量資料探勘 (Big Data Mining) 的基礎概念及應用技術。
- 課程內容包括
  - 巨量資料探勘 (Big Data Mining)
  - 大數據、AI人工智慧與深度學習 (Big Data, Artificial Intelligence and Deep Learning)
  - 關連分析 (Association Analysis)
  - 分類與預測 (Classification and Prediction)
  - 分群分析 (Cluster Analysis)
  - SAS企業資料採礦實務 (SAS Enterprise Miner)
  - 巨量資料探勘個案分析與實作

Course Introduction

- This course introduces the fundamental concepts and applications technology of big data mining.
- Topics include
  - Big Data Mining
  - Big Data, Artificial Intelligence and Deep Learning
  - Association Analysis
  - Classification and Prediction
  - Cluster Analysis
  - Data Mining Using SAS Enterprise Miner (SAS EM)
  - Case Study and Implementation of Big Data Mining

課程目標
(Objective)

- 瞭解及應用巨量資料探勘基本概念與技術。

- Understand and apply the fundamental concepts and technology of big data mining

課程大綱 (Syllabus)

週次 (Week)	日期 (Date)	內容(Subject/Topics)
1	2018/02/28	和平紀念日(放假一天) (Peace Memorial Day) (Day off)
2	2018/03/07	巨量資料探勘課程介紹 (Course Orientation for Big Data Mining)
3	2018/03/14	大數據、AI人工智慧與深度學習 (Big Data, Artificial Intelligence and Deep Learning)
4	2018/03/21	關連分析 (Association Analysis)
5	2018/03/28	分類與預測 (Classification and Prediction)
6	2018/04/04	兒童節(放假一天)(Children's Day) (Day off)
7	2018/04/11	分群分析 (Cluster Analysis)
8	2018/04/18	個案分析與實作一 (SAS EM 分群分析)：Case Study 1 (Cluster Analysis - K-Means using SAS EM)

課程大綱 (Syllabus)

週次 (Week)	日期 (Date)	內容 (Subject/Topics)
9	2018/04/25	期中報告 (Midterm Project Presentation)
10	2018/05/02	期中考試週
11	2018/05/09	個案分析與實作二 (SAS EM 關連分析)： Case Study 2 (Association Analysis using SAS EM)
12	2018/05/16	個案分析與實作三 (SAS EM 決策樹、模型評估)： Case Study 3 (Decision Tree, Model Evaluation using SAS EM)
13	2018/05/23	個案分析與實作四 (SAS EM 迴歸分析、類神經網路)： Case Study 4 (Regression Analysis, Artificial Neural Network using SAS EM)
14	2018/05/30	期末報告 (Final Project Presentation)
15	2018/06/06	畢業考試週

教學方法與評量方法

- 教學方法
  - 講述、討論、賞析、模擬、實作、問題解決
- 評量方法
  - 紙筆測驗、實作、報告、上課表現

教材課本

- 教材課本
  - 講義 (Slides)
  - 資料採礦運用：以SAS Enterprise Miner為工具，李淑娟，2015，SAS賽仕電腦軟體
- 參考書籍
  - Big Data, Data Mining, and Machine Learning: Value Creation for Business Leaders and Practitioners, Jared Dean, Wiley, 2014
  - Data Science for Business: What you need to know about data mining and data-analytic thinking, Foster Provost and Tom Fawcett, O'Reilly, 2013
  - Applied Analytics Using SAS Enterprise Mining, Jim Georges, Jeff Thompson and Chip Wells, SAS, 2010
  - Data Mining: Concepts and Techniques, Third Edition, Jiawei Han, Micheline Kamber and Jian Pei, Morgan Kaufmann, 2011

作業與學期成績計算方式

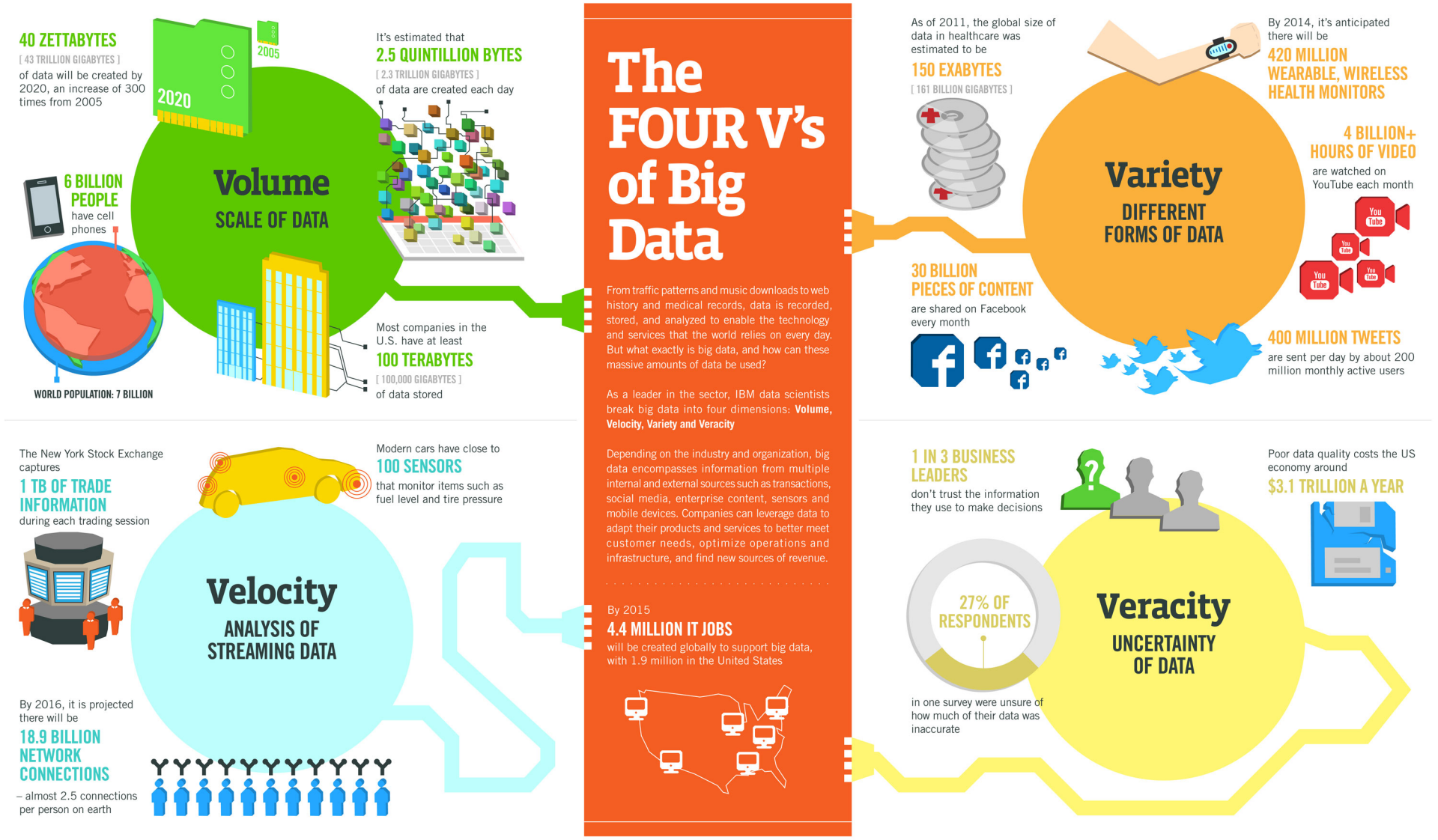
- 作業篇數
  - 3篇
- 學期成績計算方式
  - ☑期中評量：30 %
  - ☑期末評量：30 %
  - ☑其他（課堂參與及報告討論表現）：40 %

Team Term Project

- Term Project Topics
  - AI Challenge
  - Big Data mining
  - Big Data Analytics
  - Business Intelligence
  - FinTech
- 3-4 人為一組
  - 分組名單於 2018/03/14 (三) 課程下課時繳交
  - 由班代統一收集協調分組名單

Big Data Analytics and Data Mining

Big Data 4 V



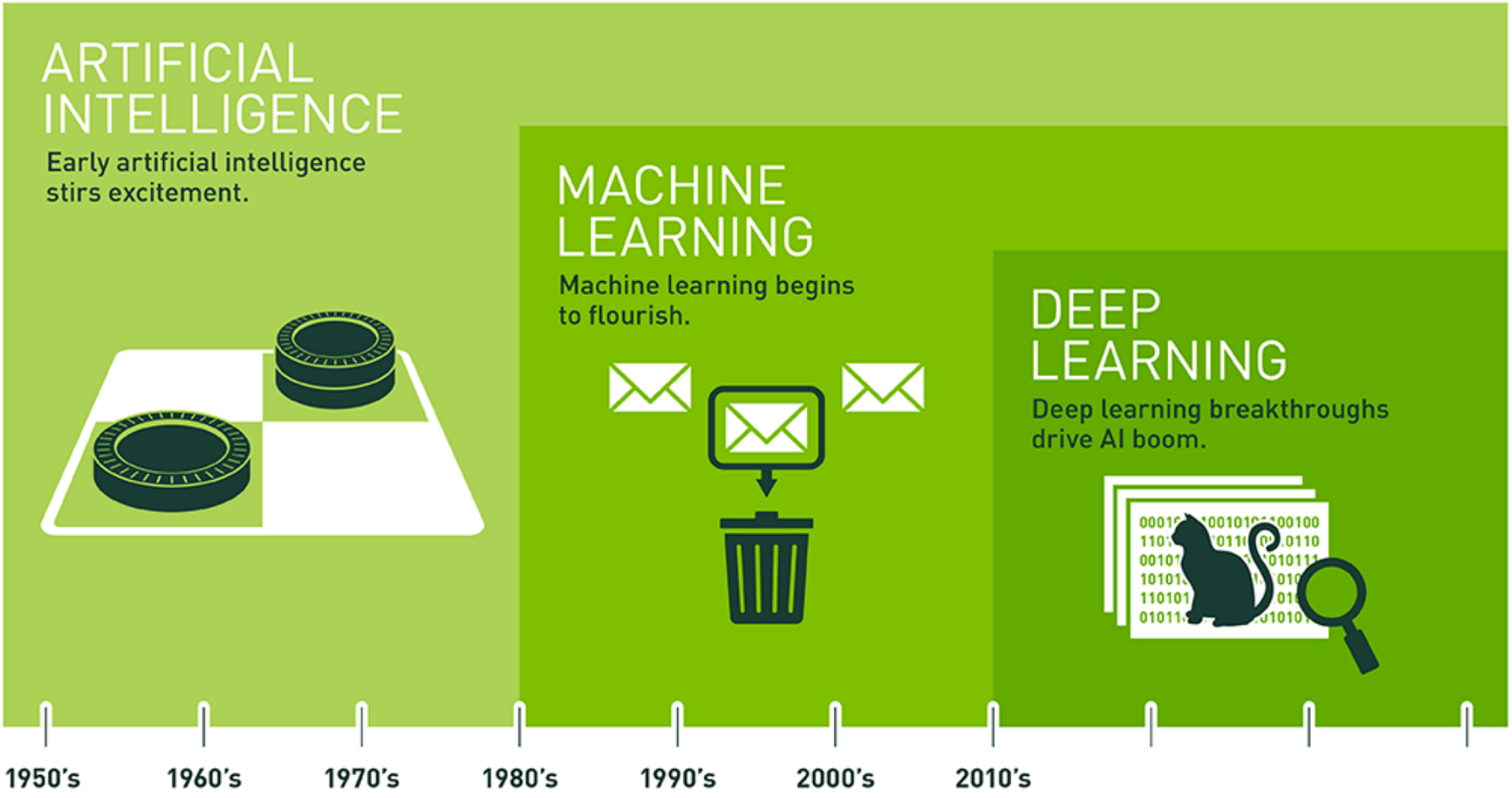
Sources: McKinsey Global Institute, Twitter, Cisco, Gartner, EMC, SAS, IBM, MEPTEC, QAS

IBM

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

Value

Artificial Intelligence
Machine Learning & Deep Learning

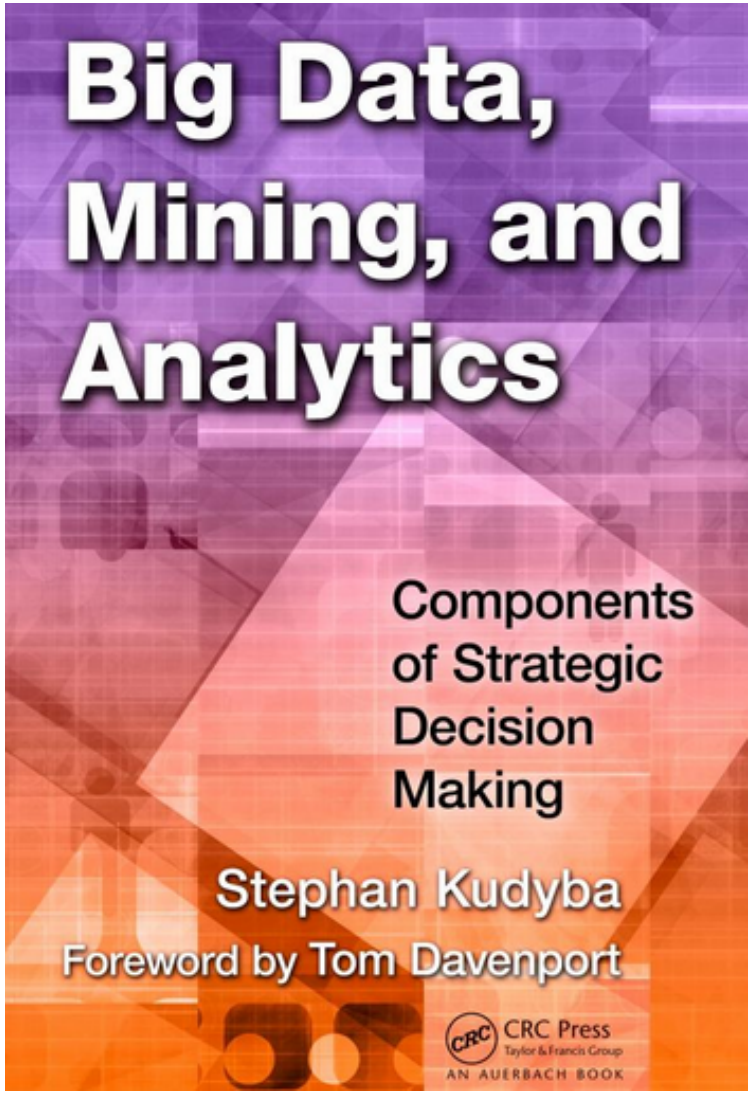


Since an early flush of optimism in the 1950s, smaller subsets of artificial intelligence – first machine learning, then deep learning, a subset of machine learning – have created ever larger disruptions.

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

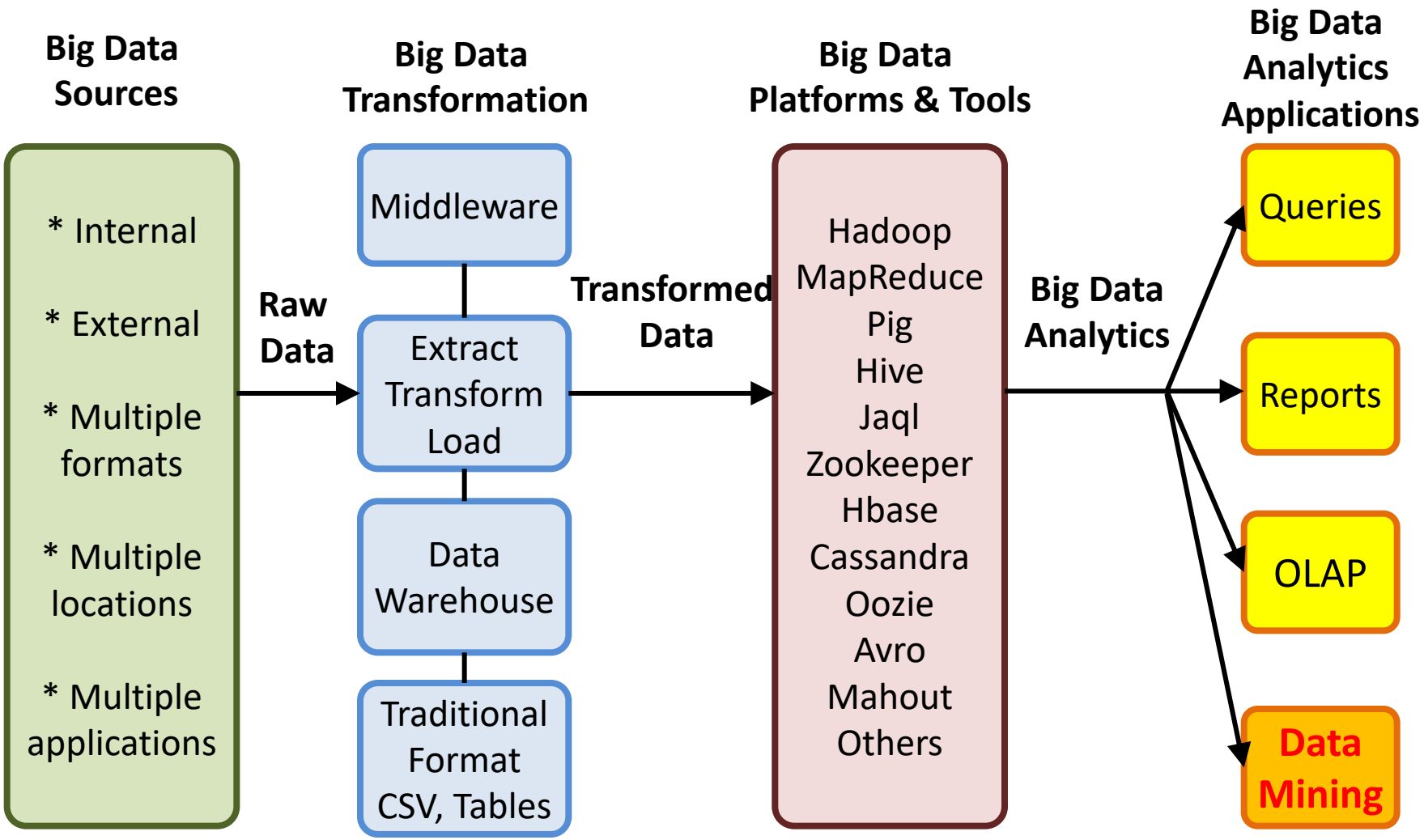
Stephan Kudyba (2014),

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



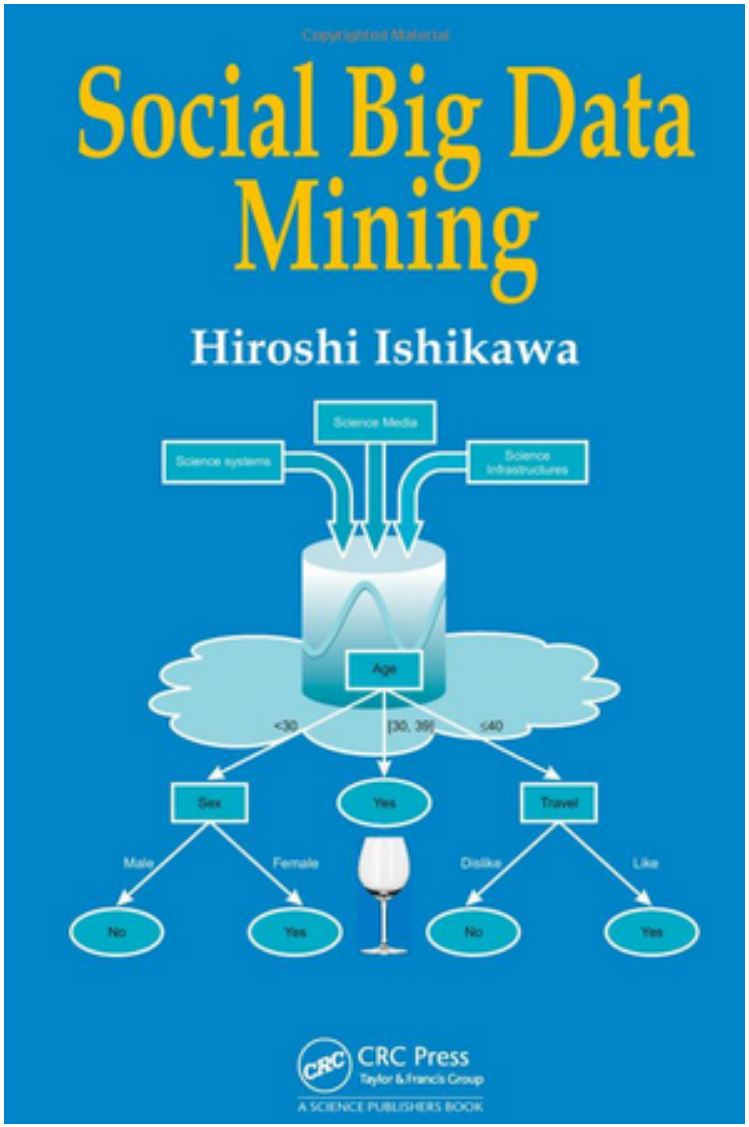
Source: Stephan Kudyba (2014), Big Data, Mining, and Analytics: Components of Strategic Decision Making, Auerbach Publications

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)



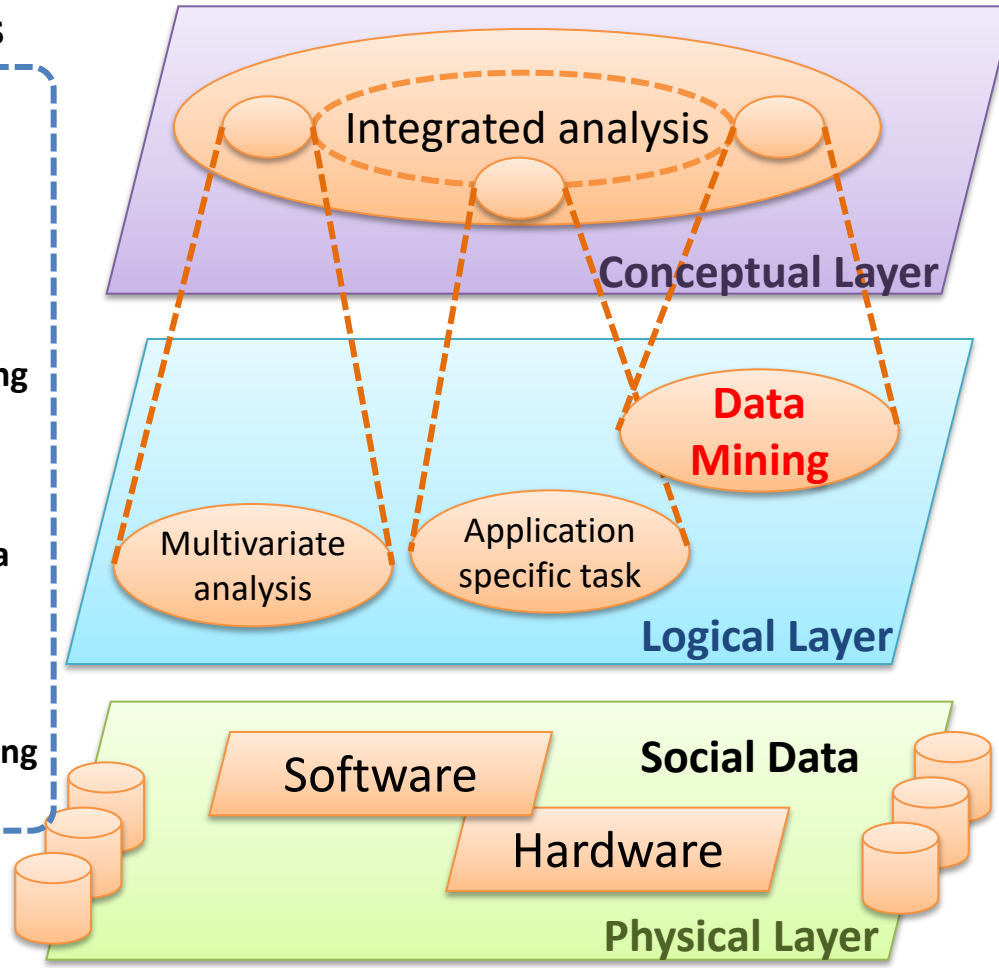
Source: http://www.amazon.com/Social-Data-Mining-Hiroshi-Ishikawa/dp/149871093X

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**

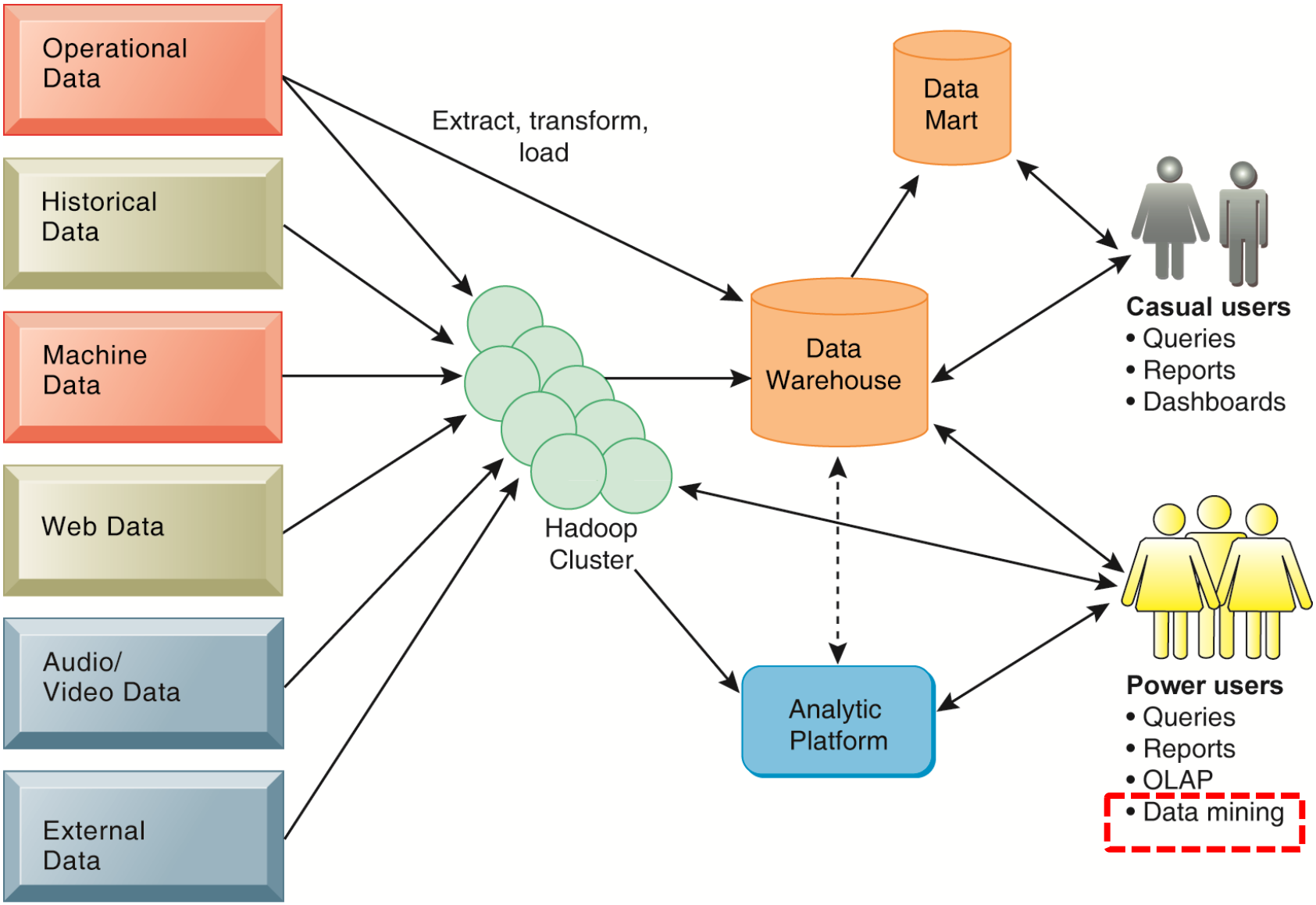


Analysts

- **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

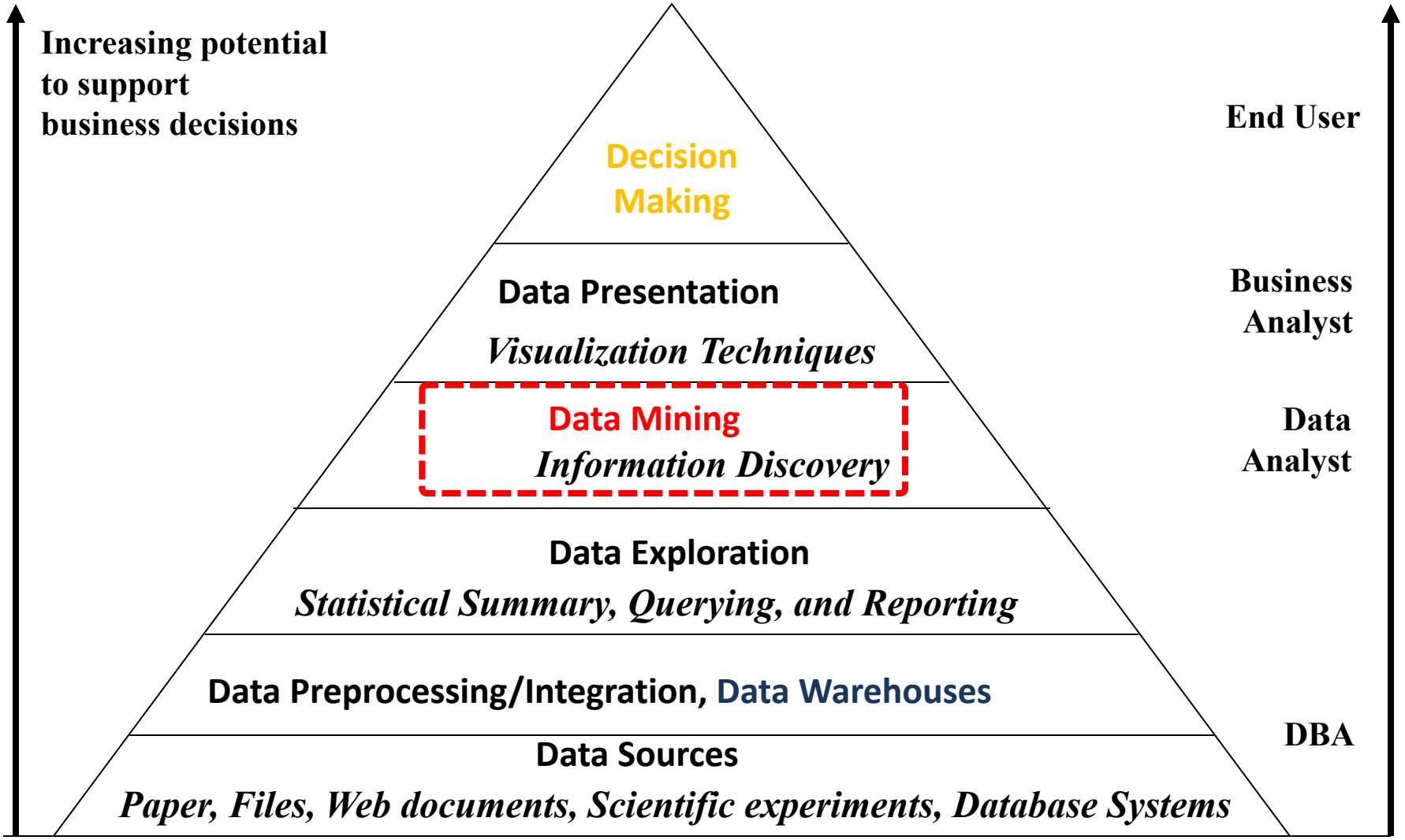
Business Intelligence (BI) Infrastructure



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

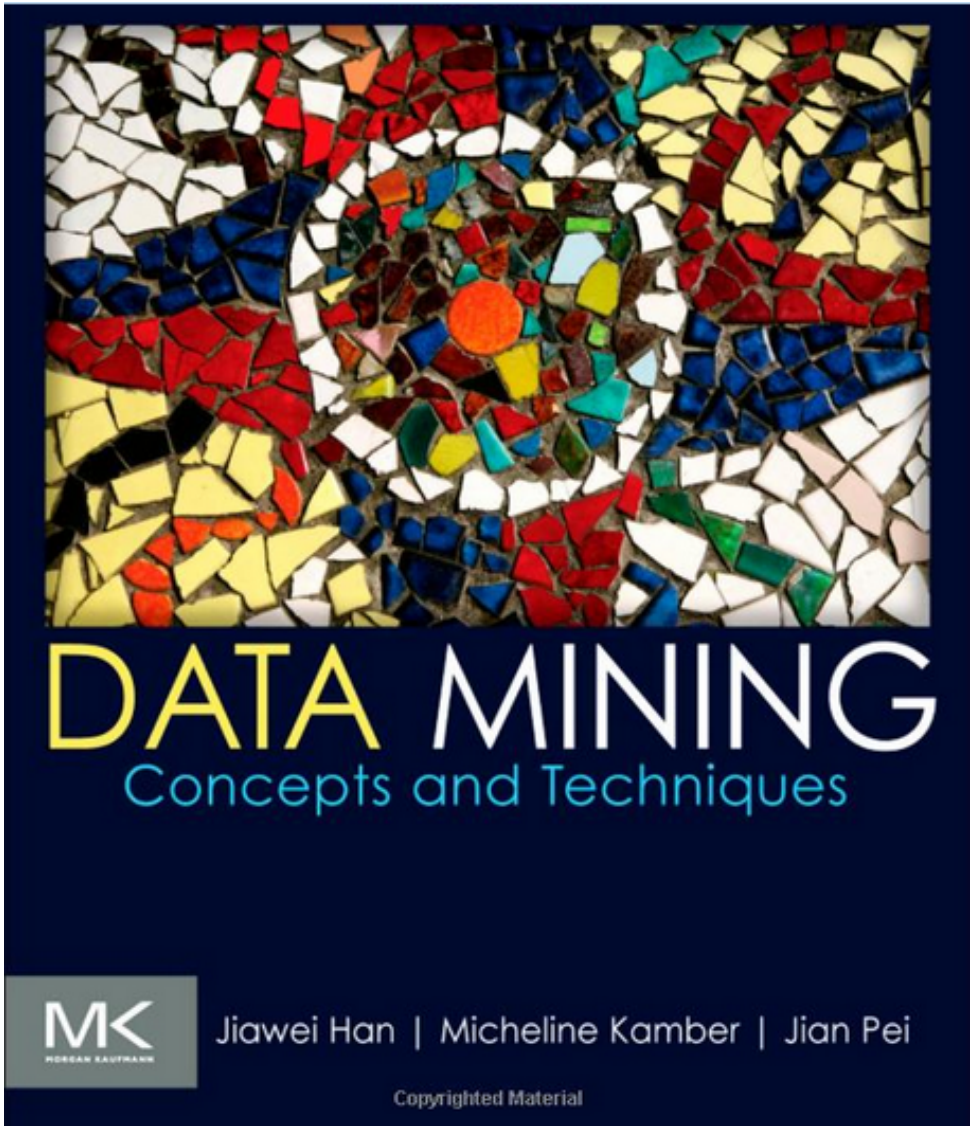


Source: Jiawei Han and Micheline Kamber (2006), Data Mining: Concepts and Techniques, Second Edition, Elsevier

The Evolution of BI Capabilities

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

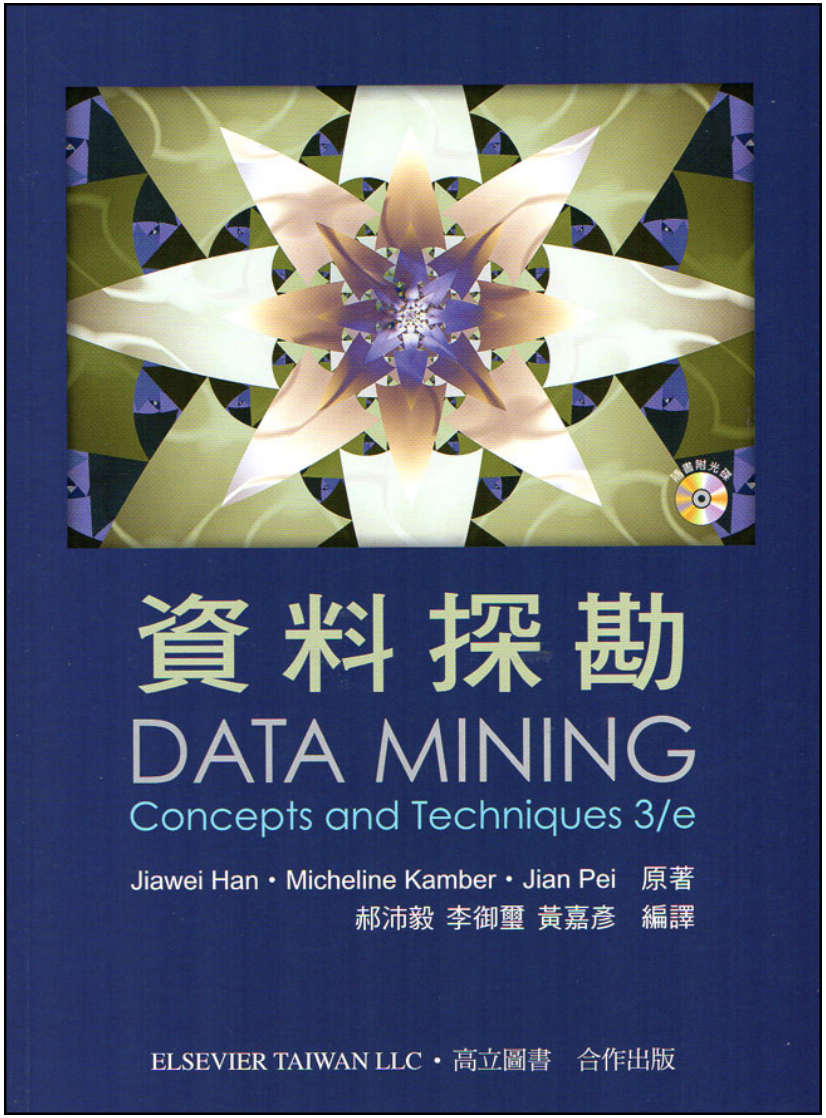
Data Mining



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

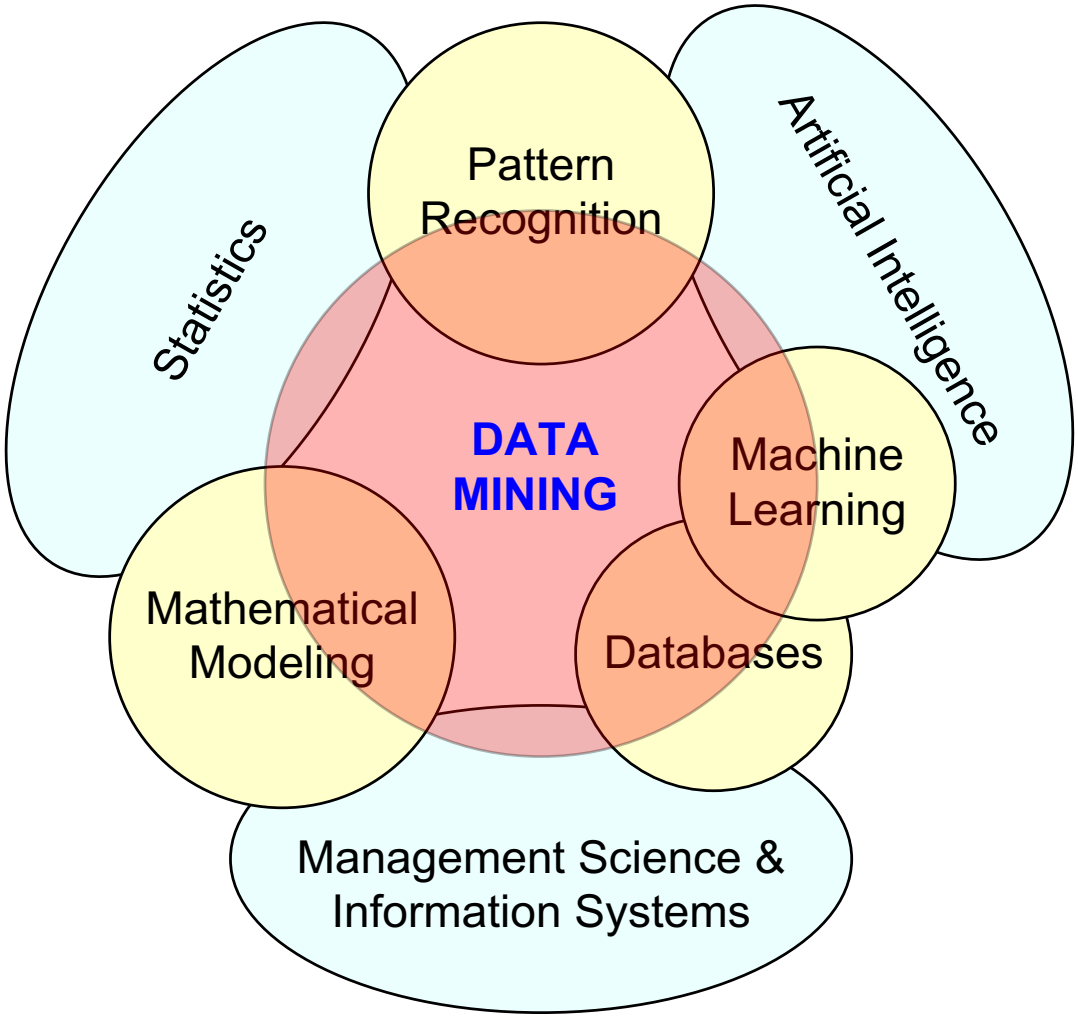
郝沛毅, 李御璽, 黃嘉彥 編譯, 資料探勘

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



Source: http://www.books.com.tw/products/0010646676

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.

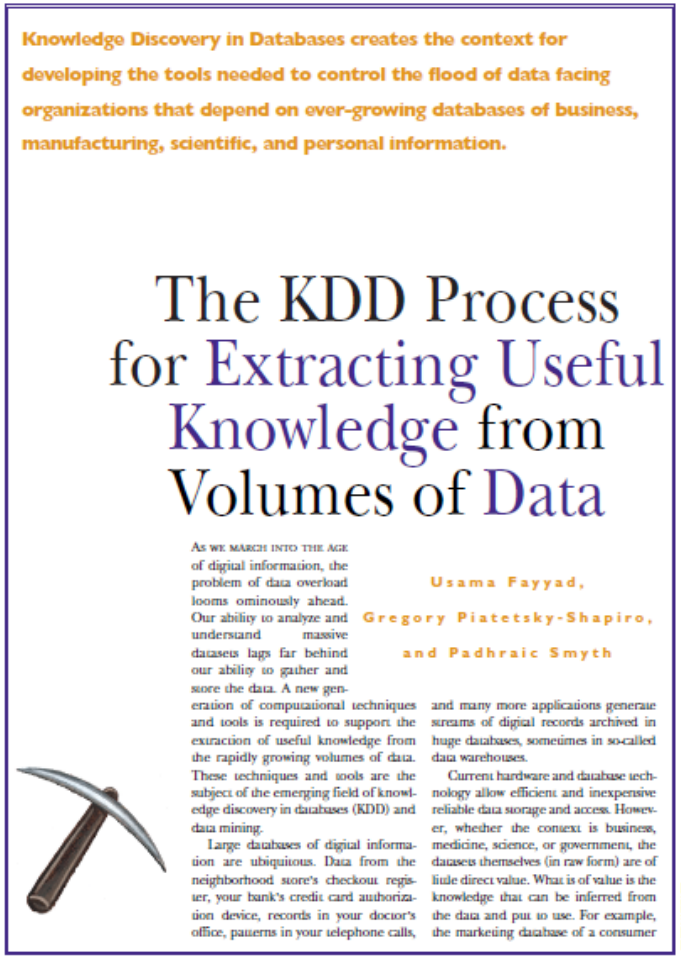
Fayyad, U., Piatetsky-Shapiro, G., & Smyth, P. (1996).

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.

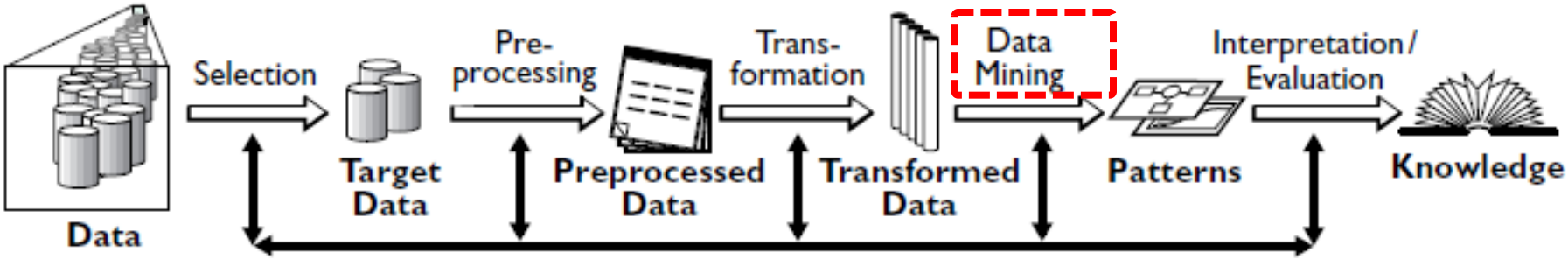
Current hardware and database technology allow efficient and inexpensive reliable data storage and access. However, whether the context is business, medicine, science, or government, the datasets themselves (in raw form) are of little direct value. What is of value is the knowledge that can be inferred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

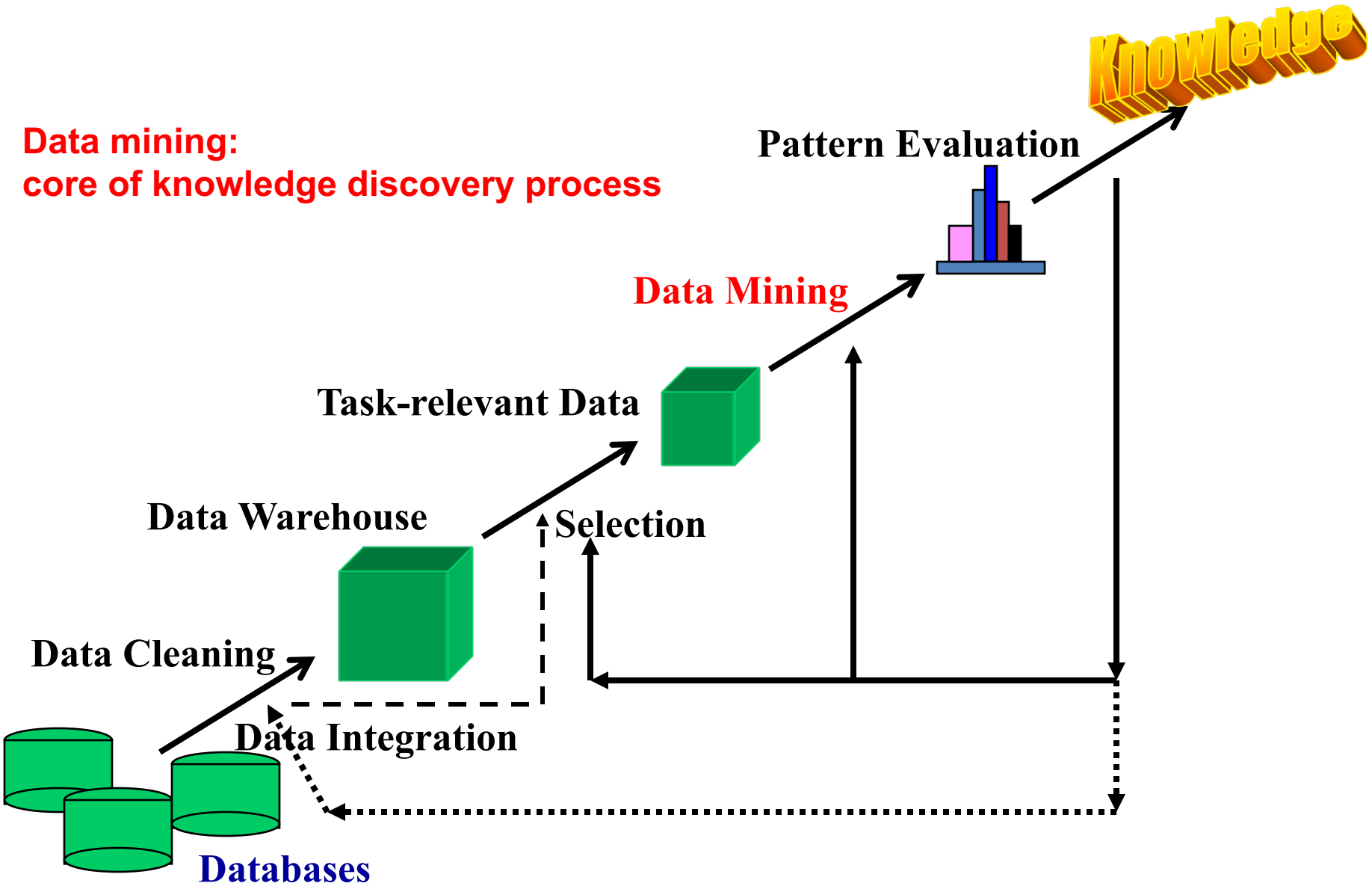
(Fayyad et al., 1996)



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

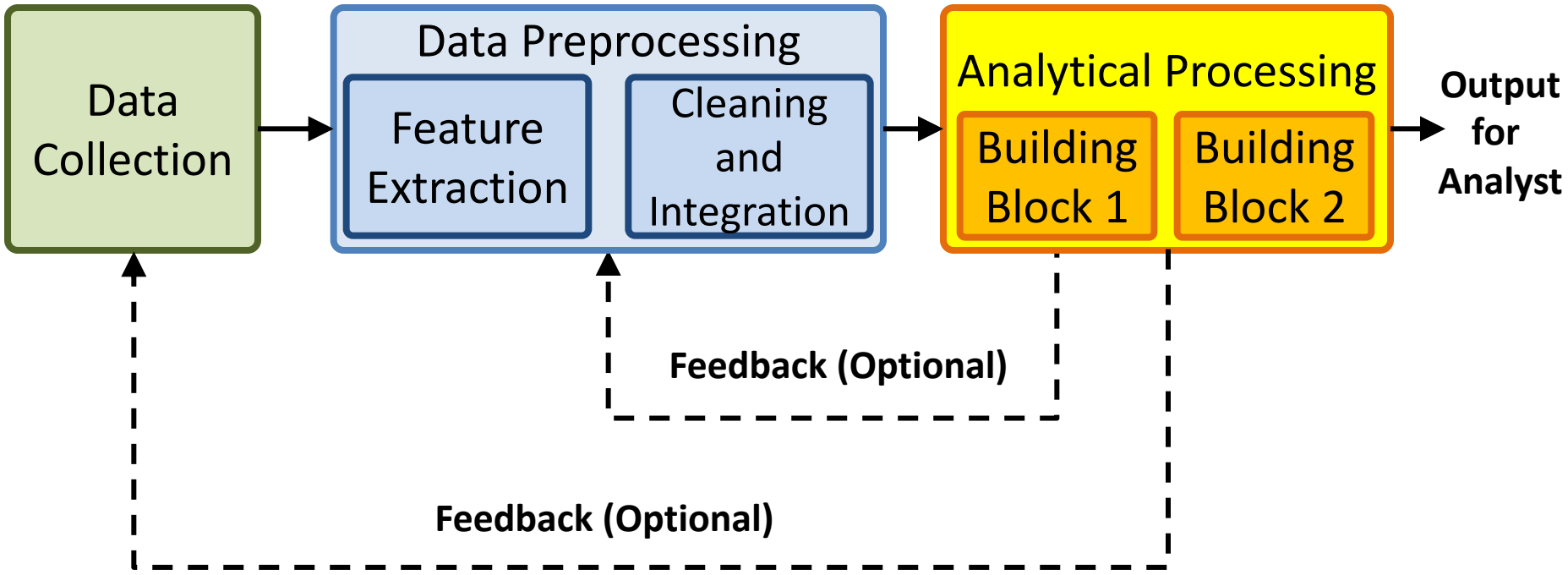
Data mining:
core of knowledge discovery process



Source: Han & Kamber (2006)

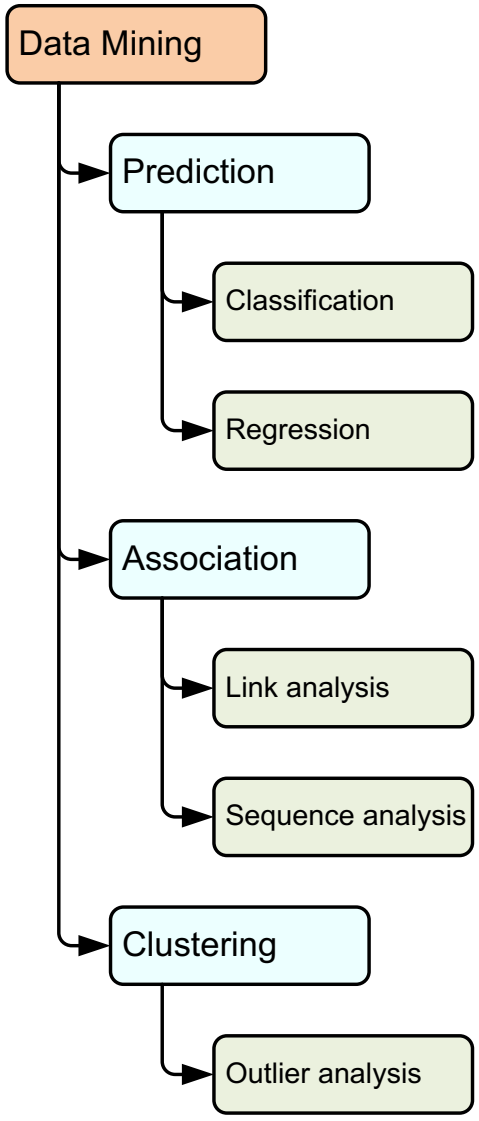
Data Mining Processing Pipeline

(Charu Aggarwal, 2015)



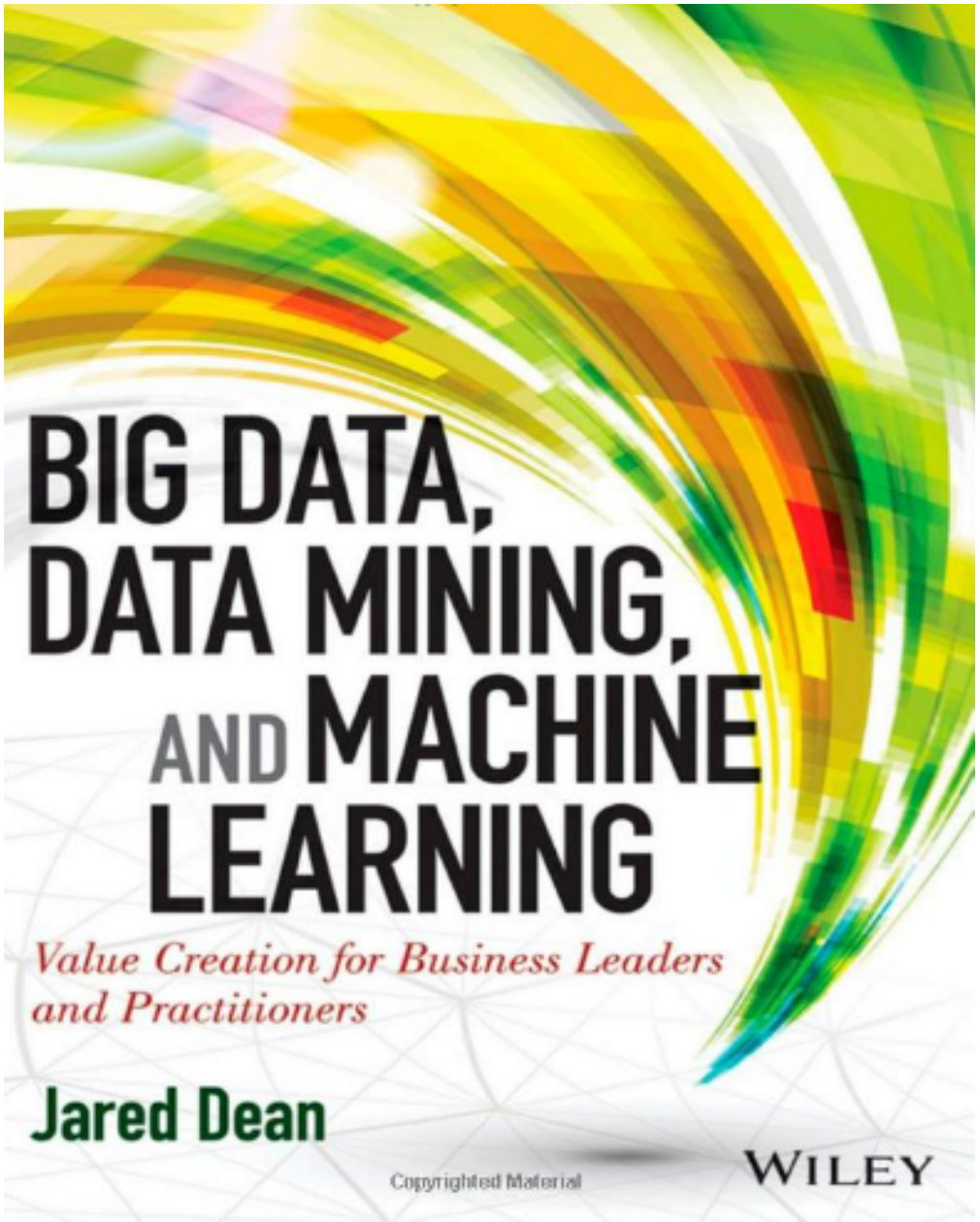
Source: Charu Aggarwal (2015), Data Mining: The Textbook Hardcover, Springer

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

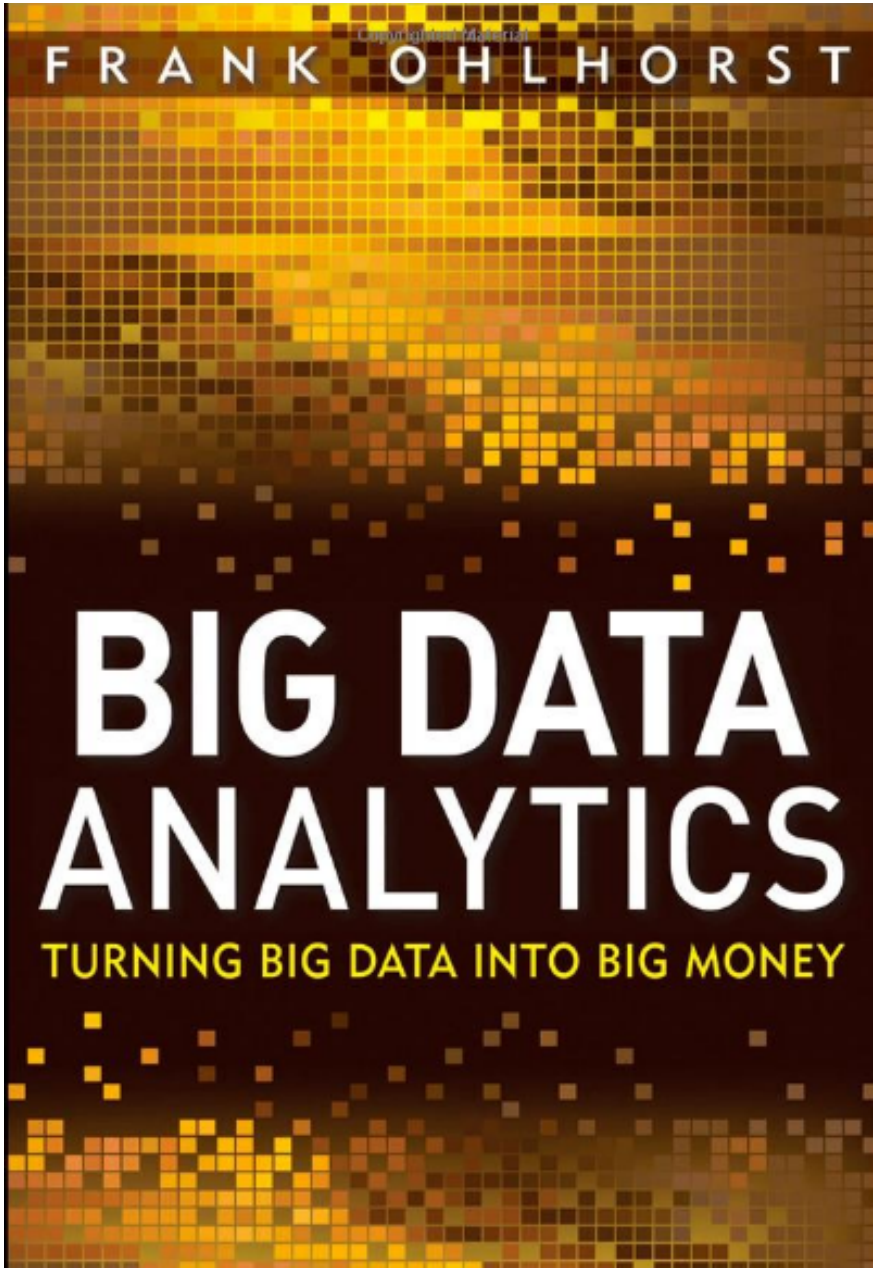


Source: http://www.amazon.com/Data-Mining-Machine-Learning-Practitioners/dp/1118618041

Deep Learning

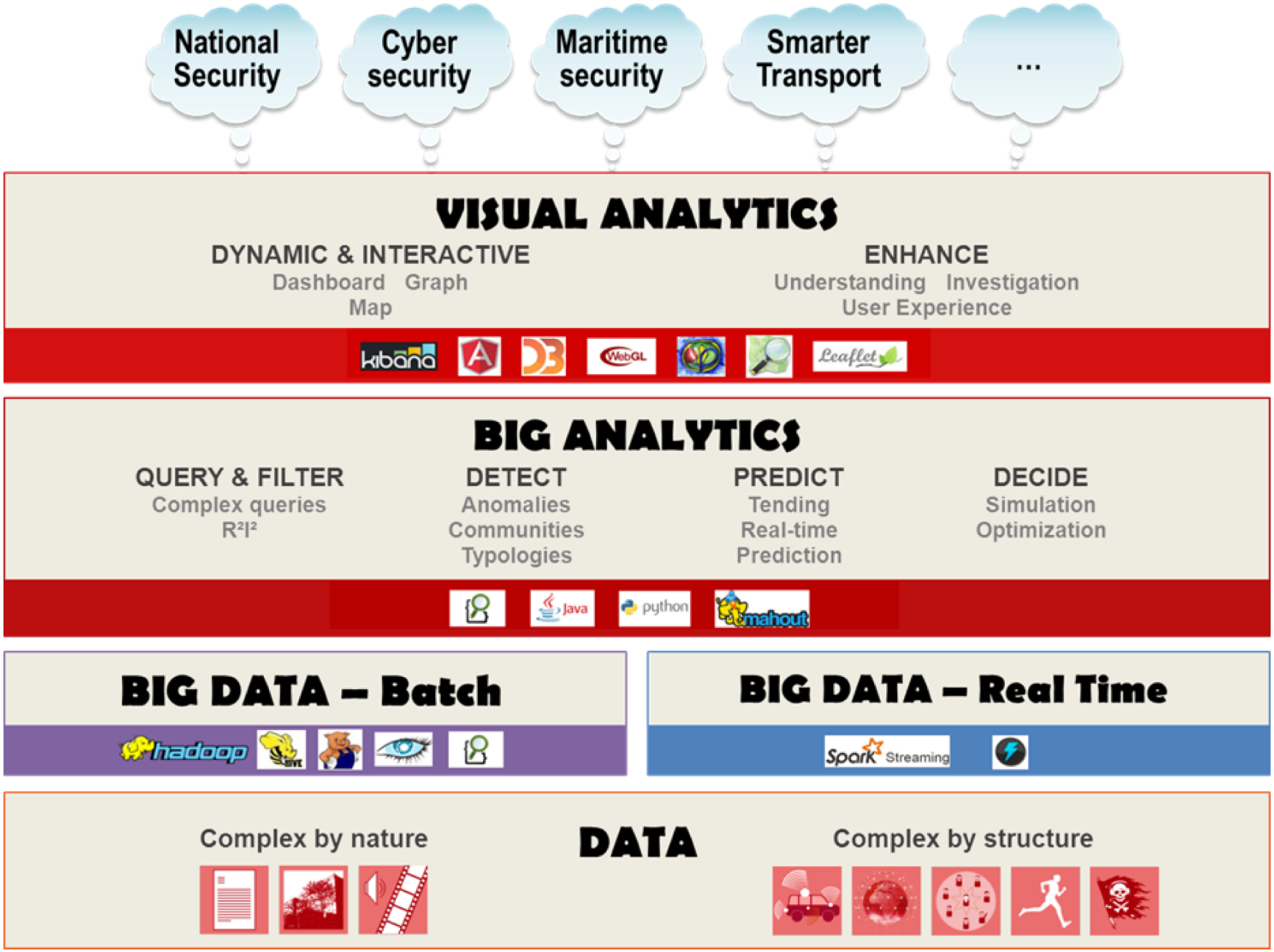
Intelligence from Big Data

Source: https://www.vlab.org/events/deep-learning/



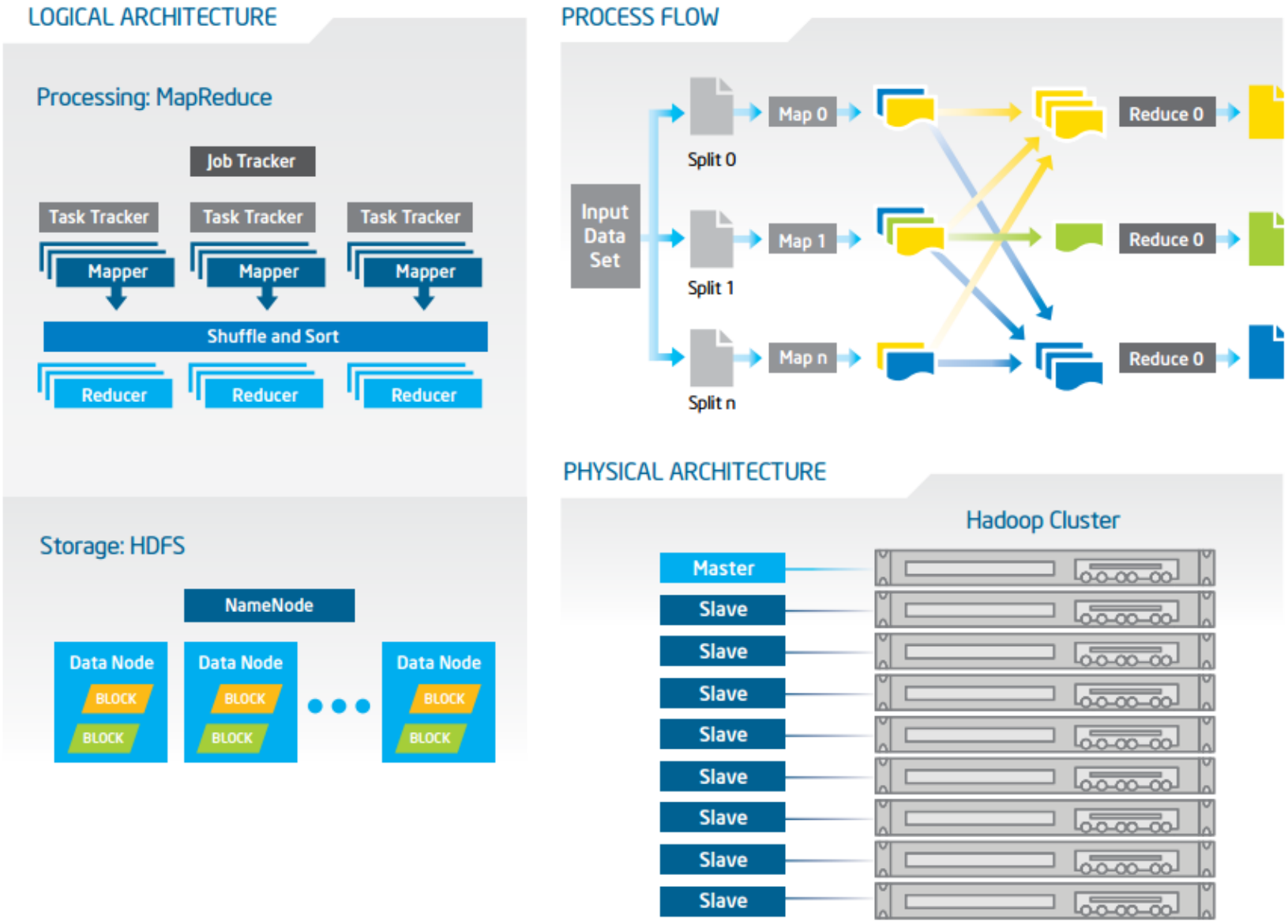
Source: http://www.amazon.com/Big-Data-Analytics-Turning-Money/dp/1118147596

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



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

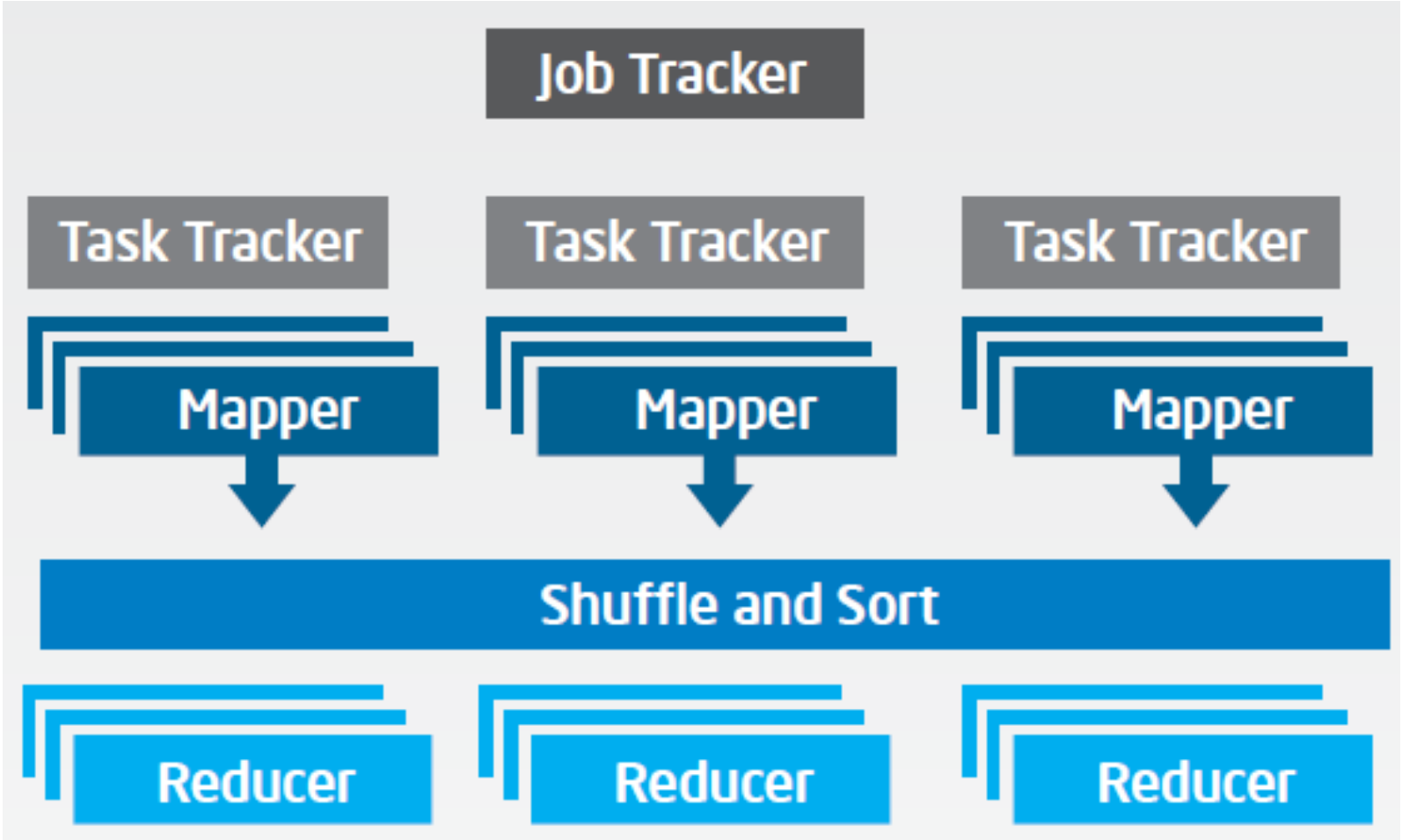


Source: https://software.intel.com/sites/default/files/article/402274/etl-big-data-with-hadoop.pdf

Big Data with Hadoop Architecture

Logical Architecture

Processing: MapReduce

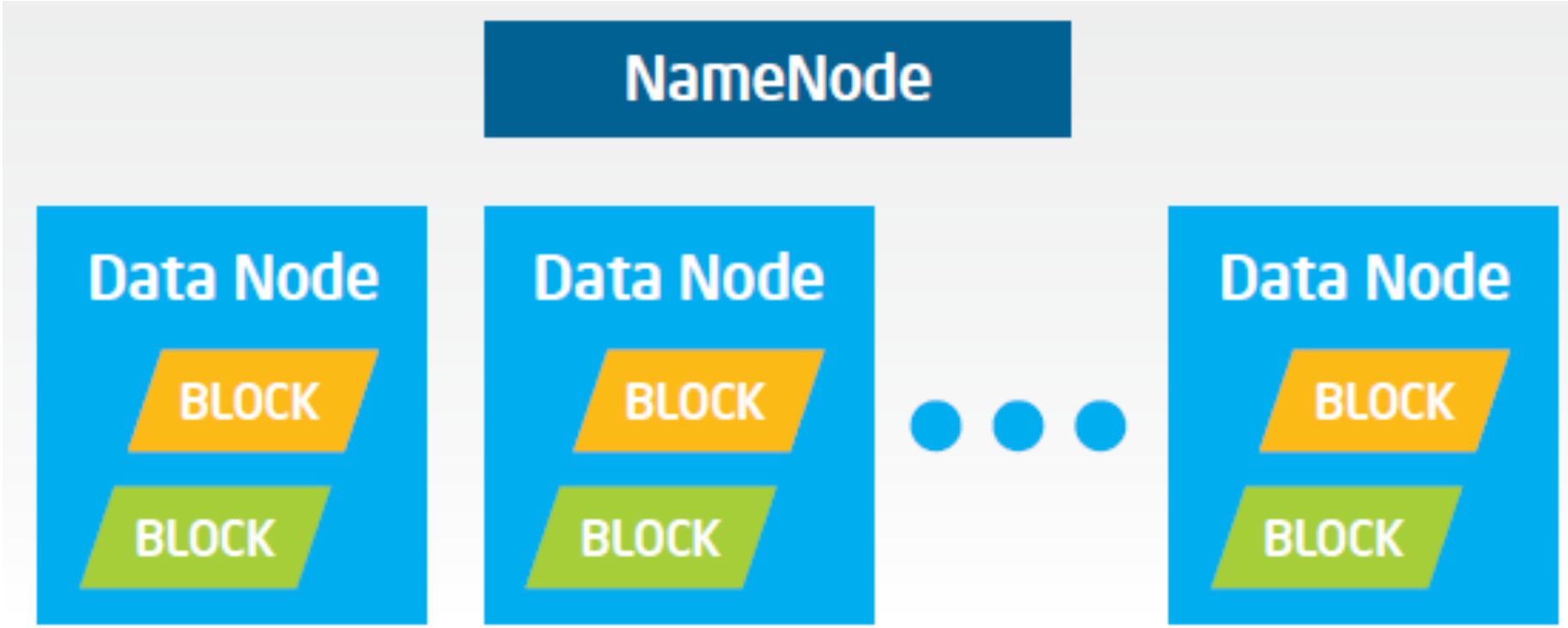


Source: https://software.intel.com/sites/default/files/article/402274/etl-big-data-with-hadoop.pdf

Big Data with Hadoop Architecture

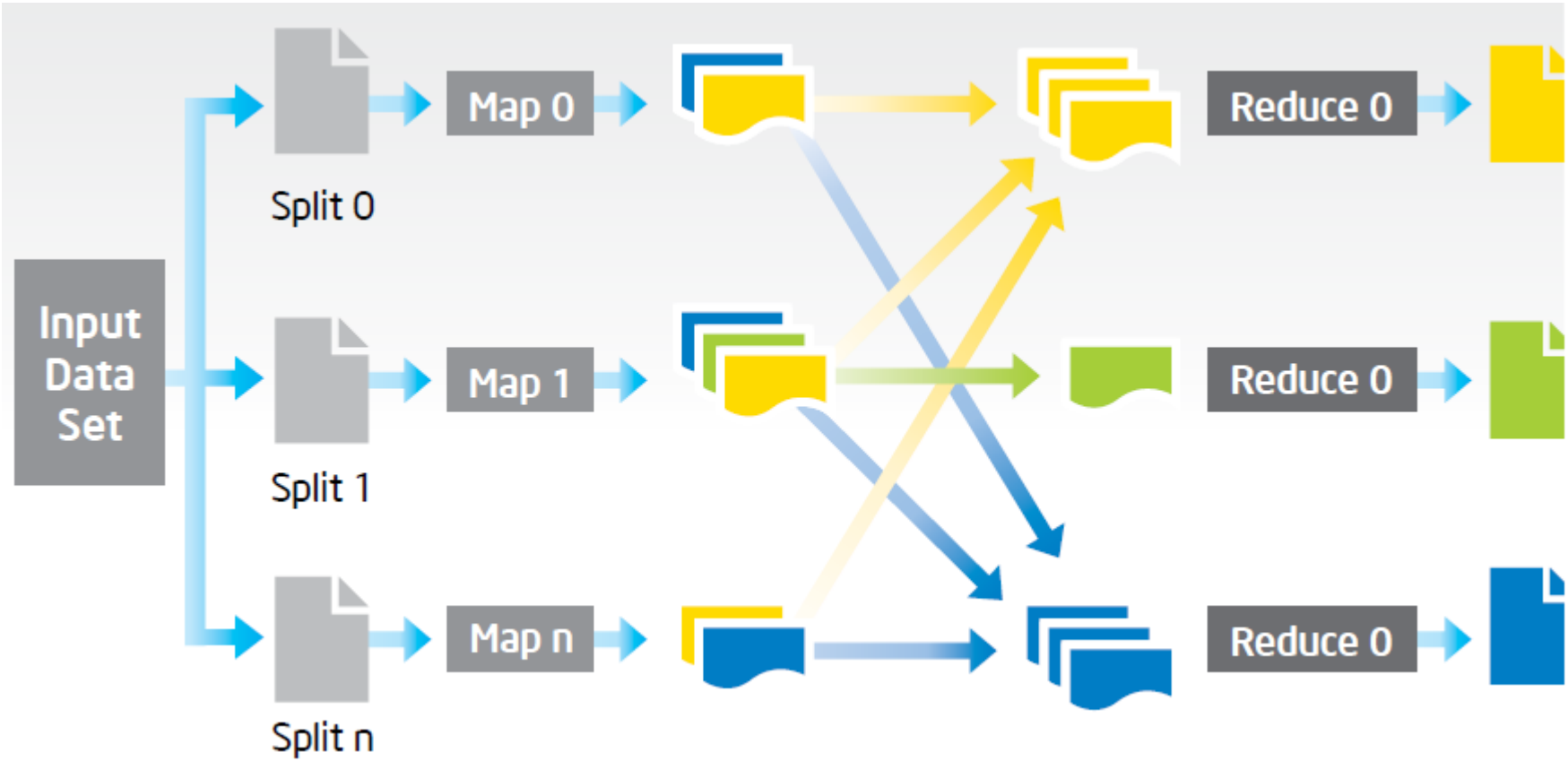
Logical Architecture

Storage: HDFS



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

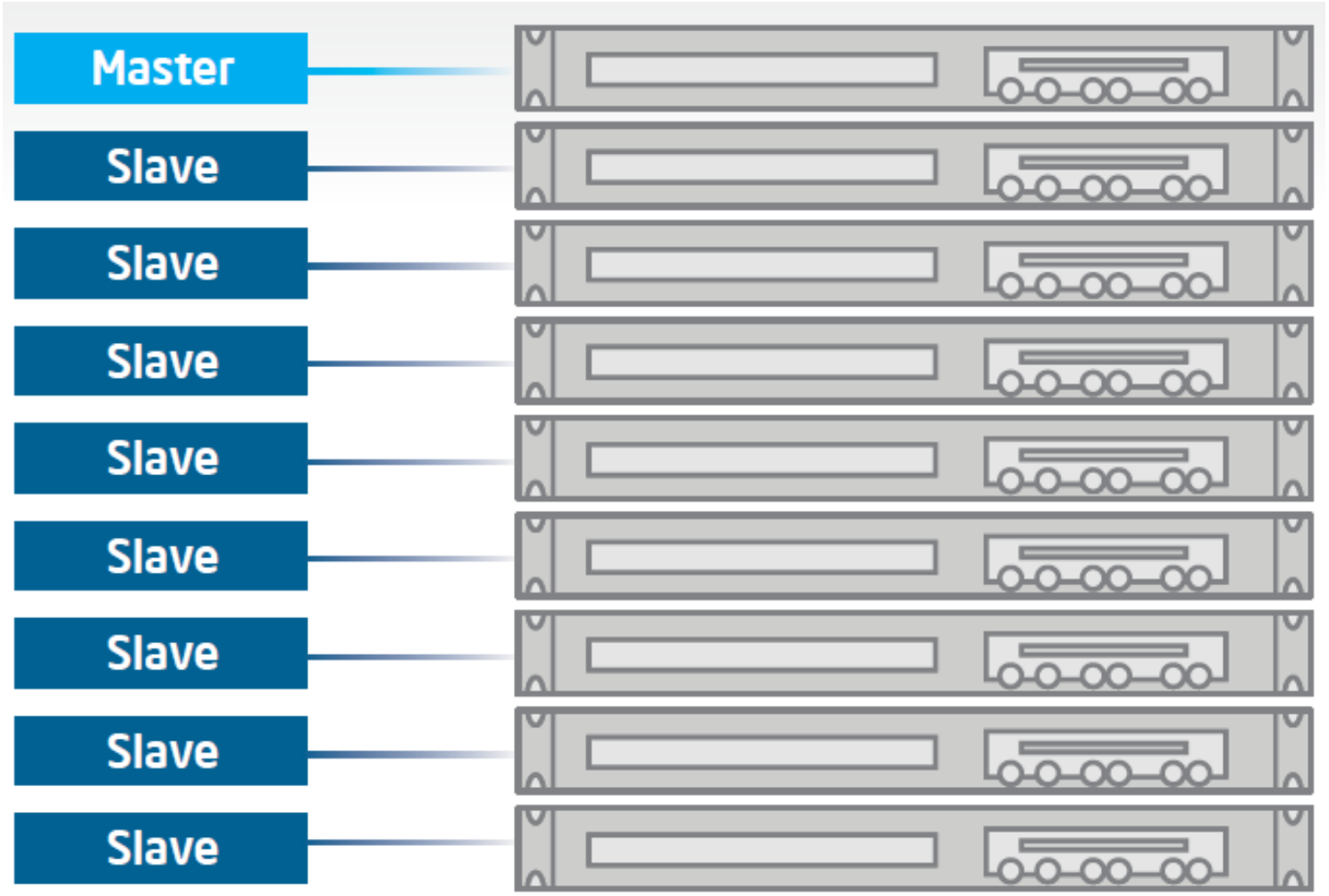
Big Data with Hadoop Architecture
Process Flow



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

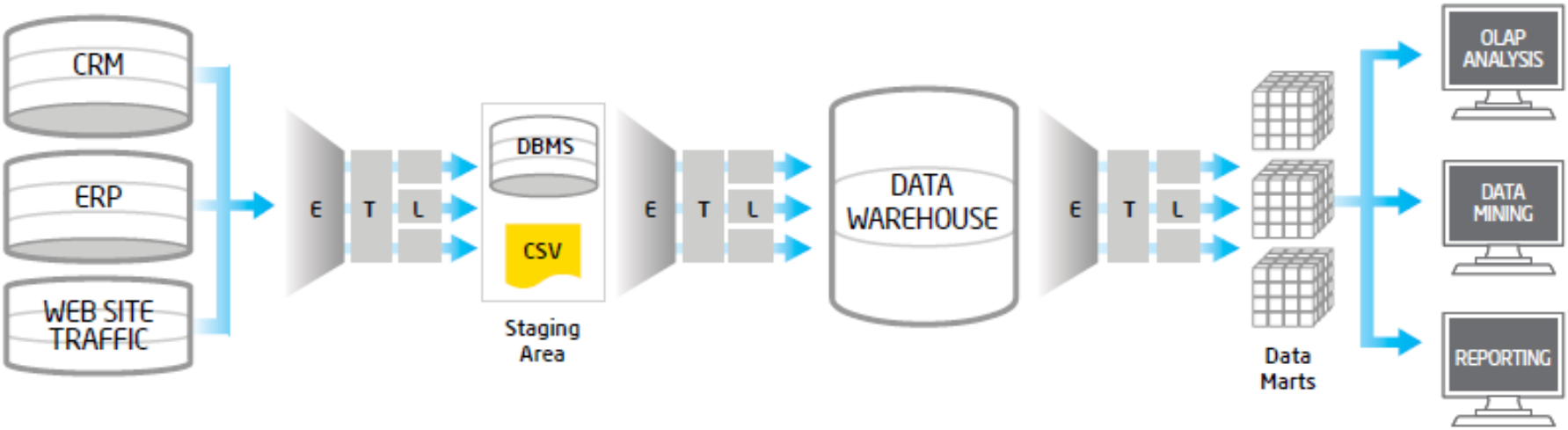
Big Data with Hadoop Architecture

Hadoop Cluster



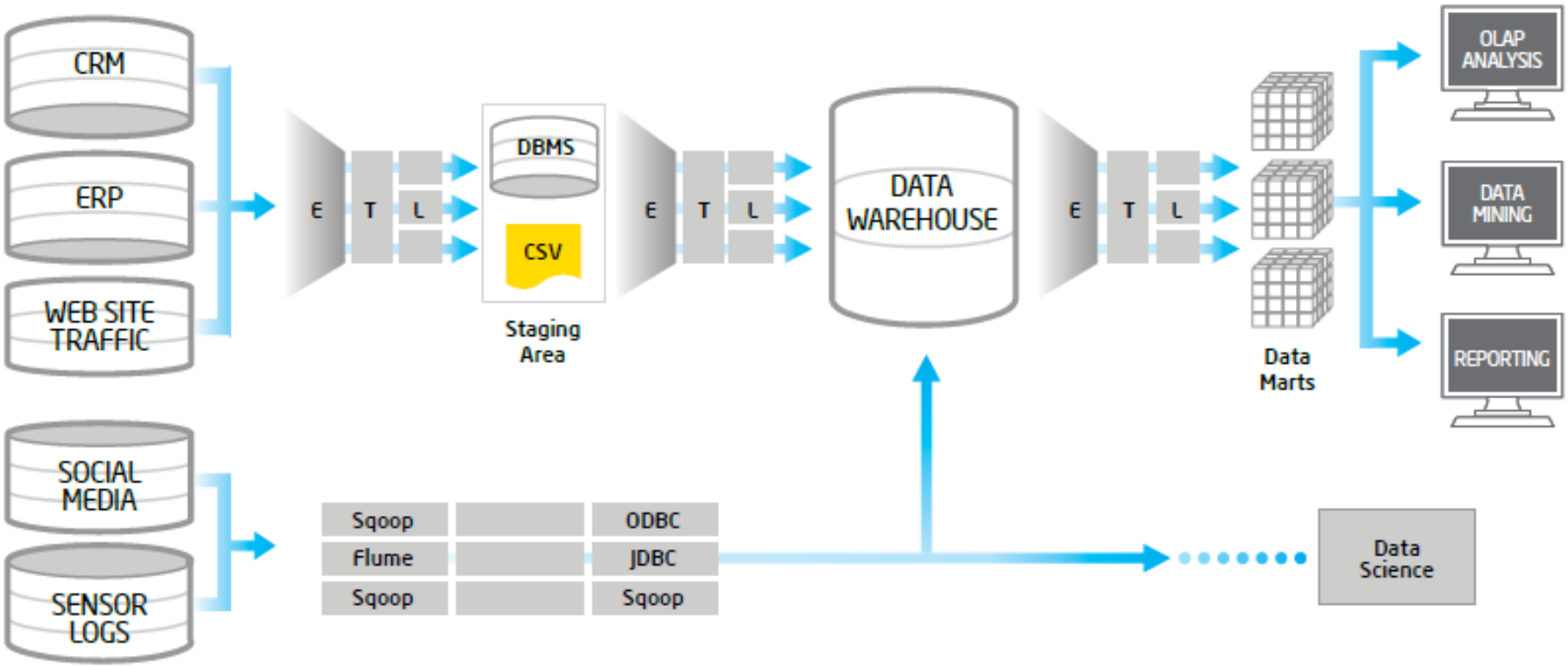
Source: https://software.intel.com/sites/default/files/article/402274/etl-big-data-with-hadoop.pdf

Traditional ETL Architecture



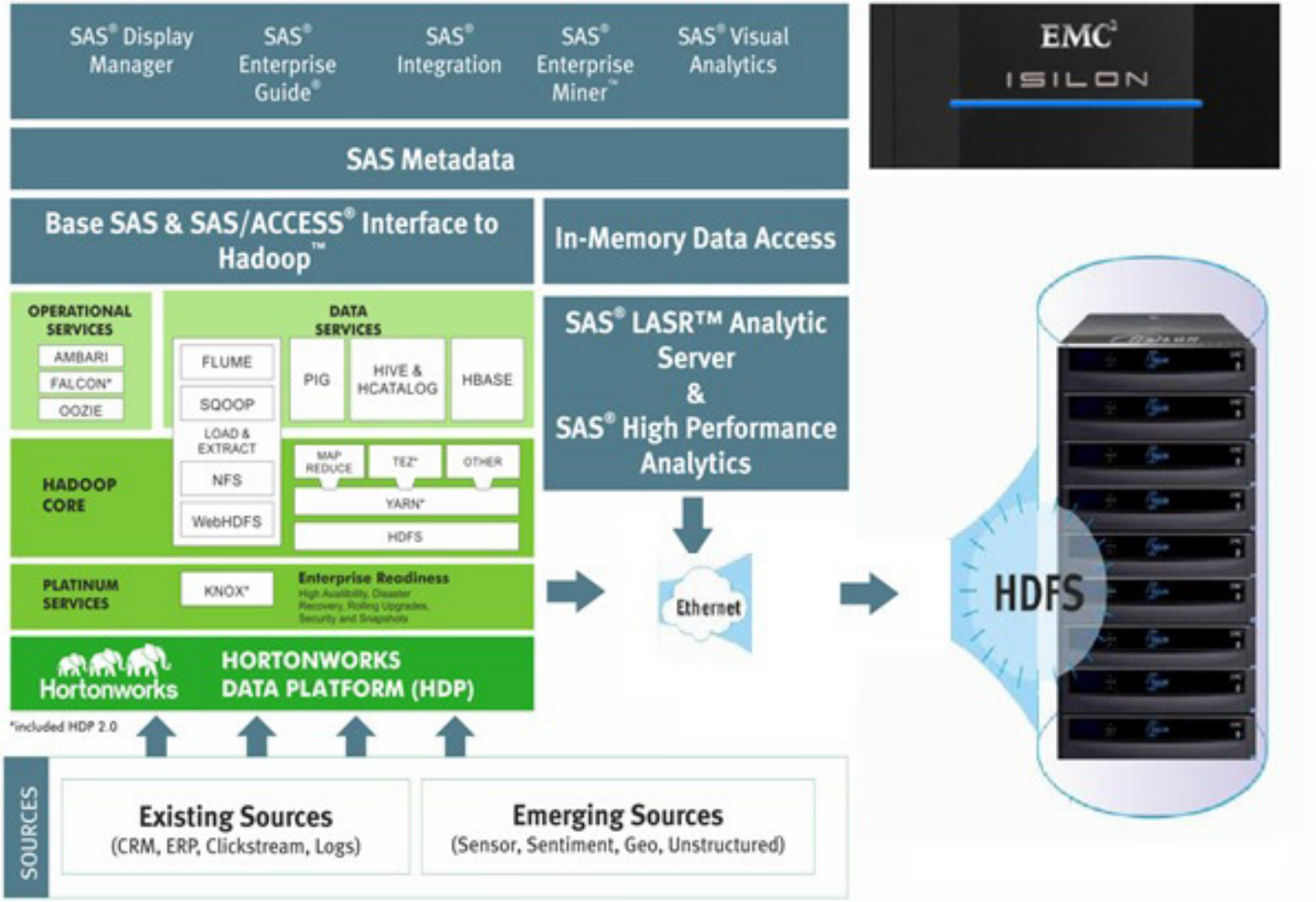
Source: https://software.intel.com/sites/default/files/article/402274/etl-big-data-with-hadoop.pdf

Offload ETL with Hadoop (Big Data Architecture)



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

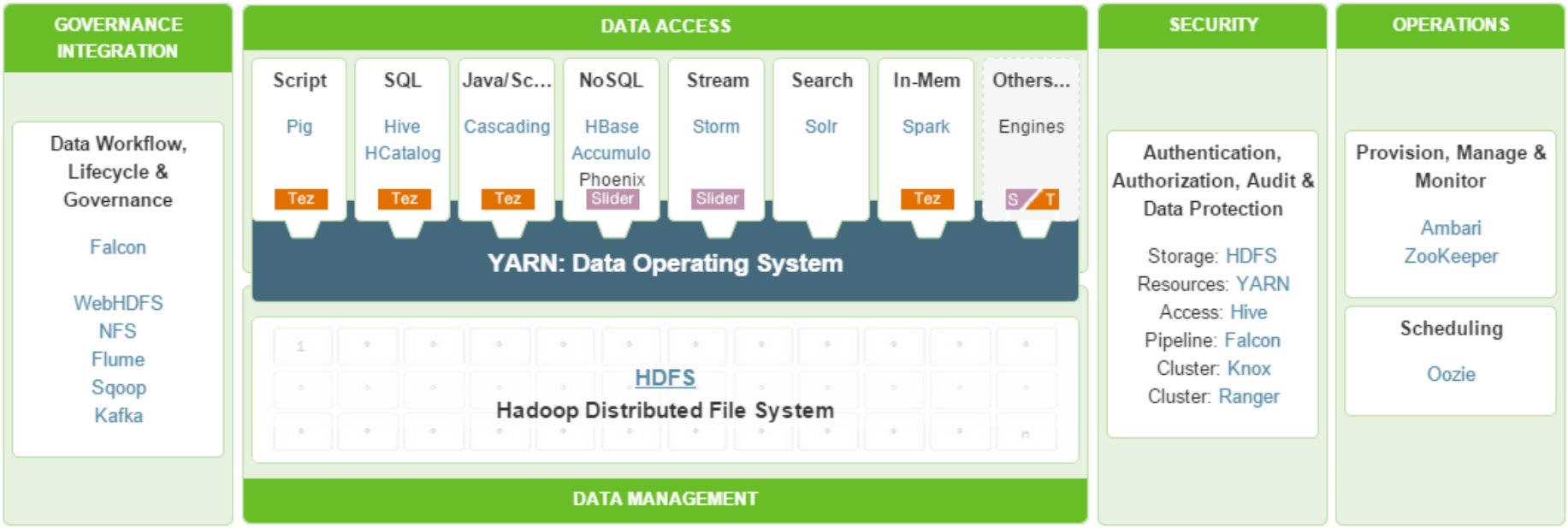
Big Data Solution



Source: http://www.newera-technologies.com/big-data-solution.html

HDP

A Complete Enterprise Hadoop Data Platform

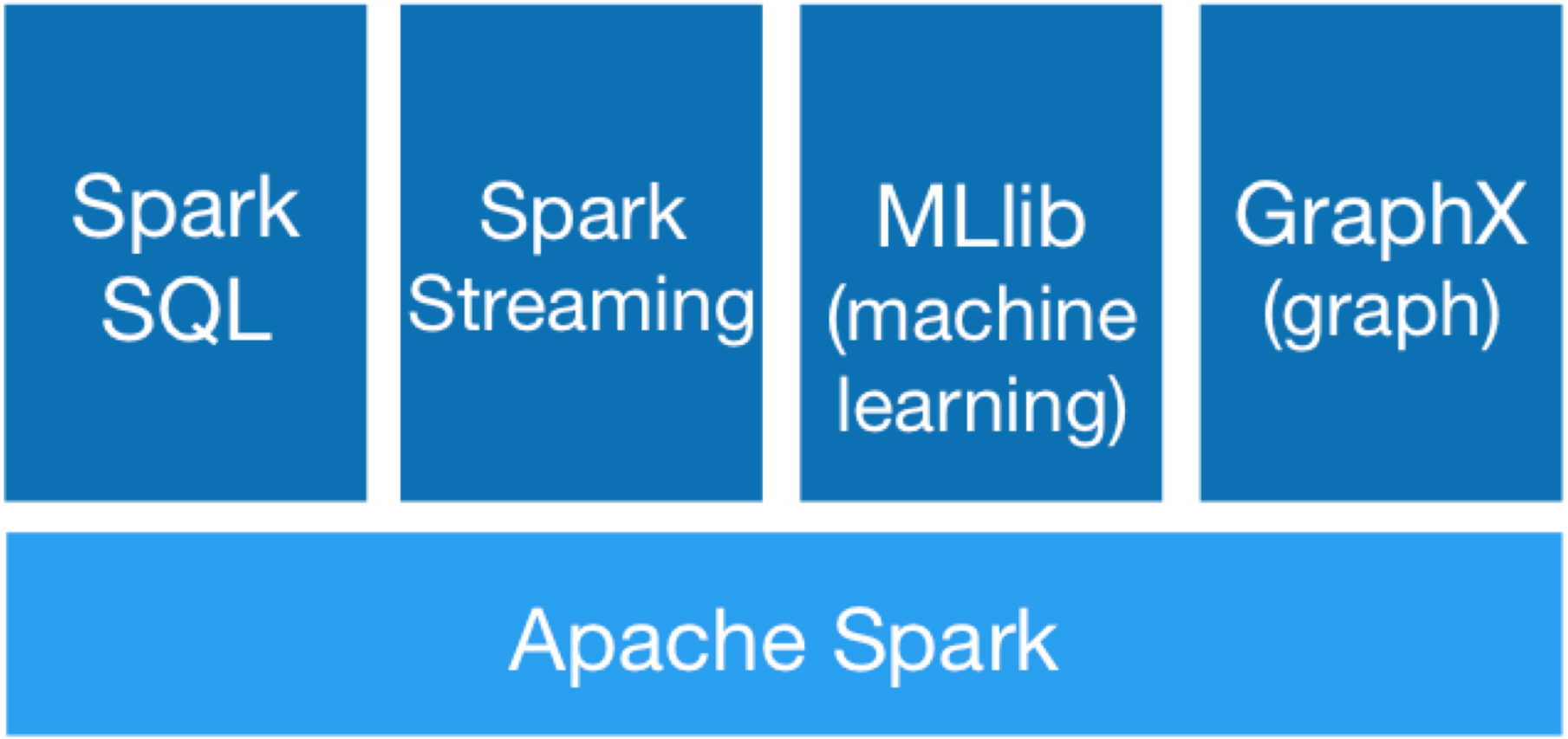


Source: http://hortonworks.com/hdp/

Spark and Hadoop

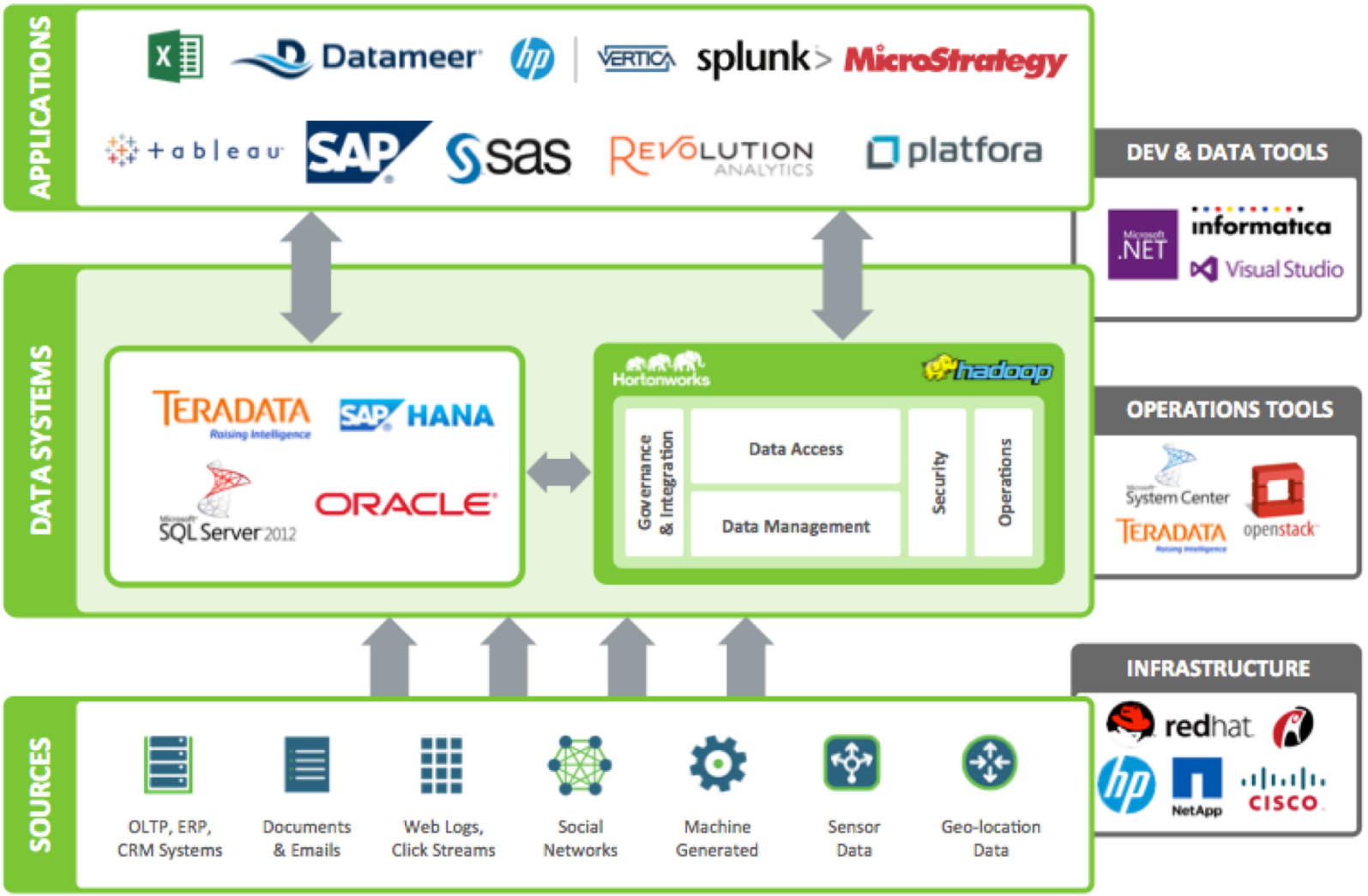
Source: http://spark.apache.org/

Spark Ecosystem



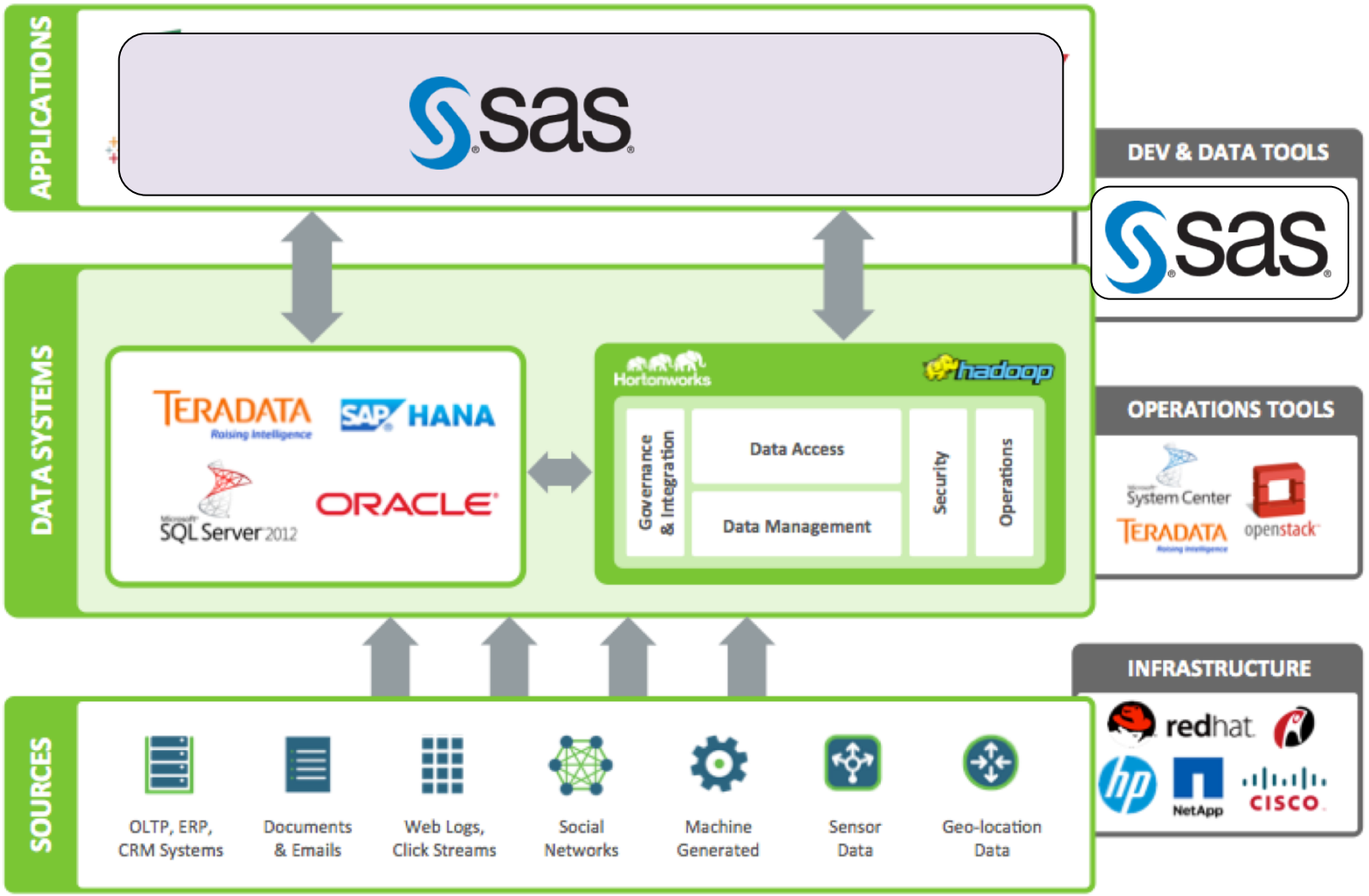
Source: http://spark.apache.org/

SAS Big data Strategy – SAS areas



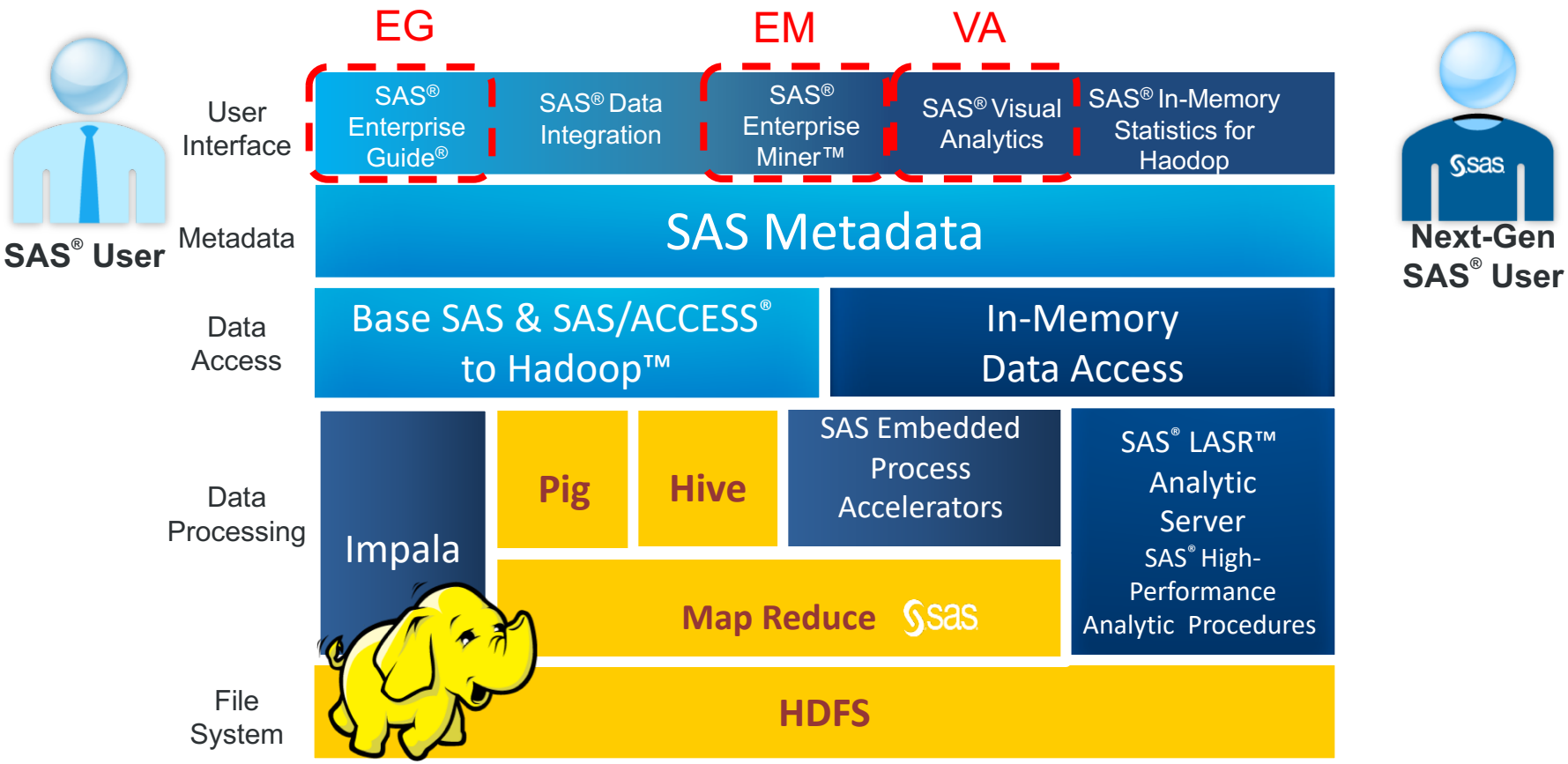
Source: Deepak Ramanathan (2014), SAS Modernization architectures - Big Data Analytics

SAS Big data Strategy – SAS areas



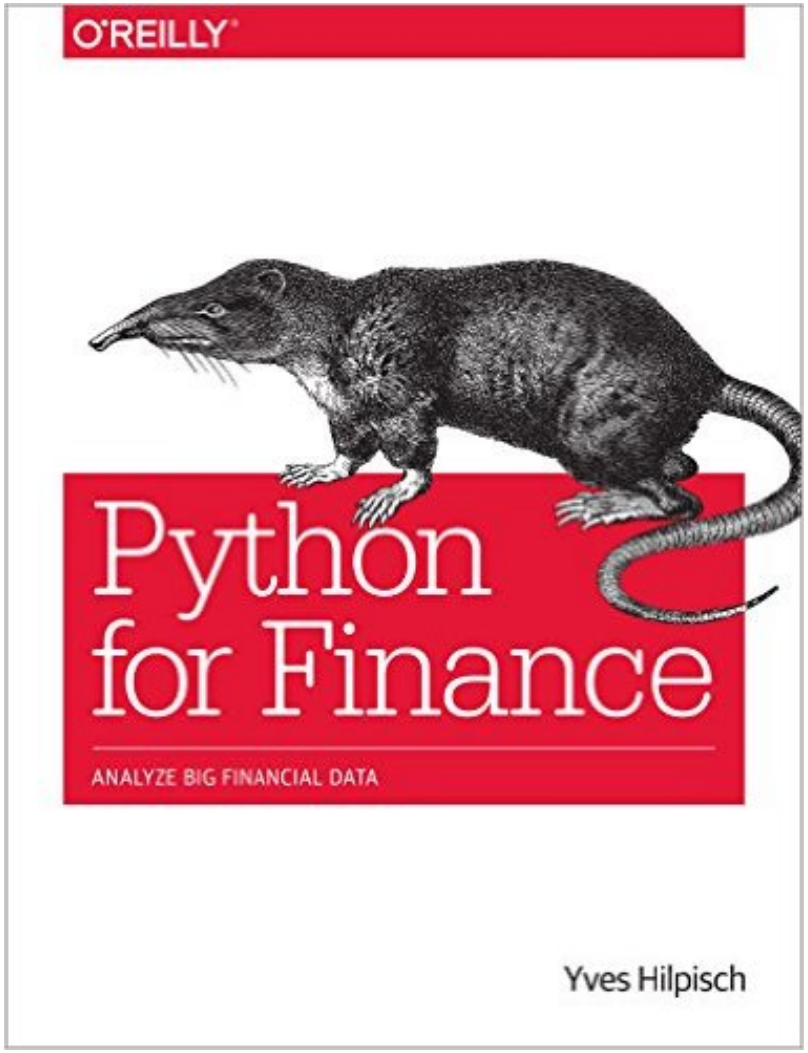
Source: Deepak Ramanathan (2014), SAS Modernization architectures - Big Data Analytics

SAS® Within the HADOOP ECOSYSTEM



Source: Deepak Ramanathan (2014), SAS Modernization architectures - Big Data Analytics

Yves Hilpisch, Python for Finance: Analyze Big Financial Data, O'Reilly, 2014

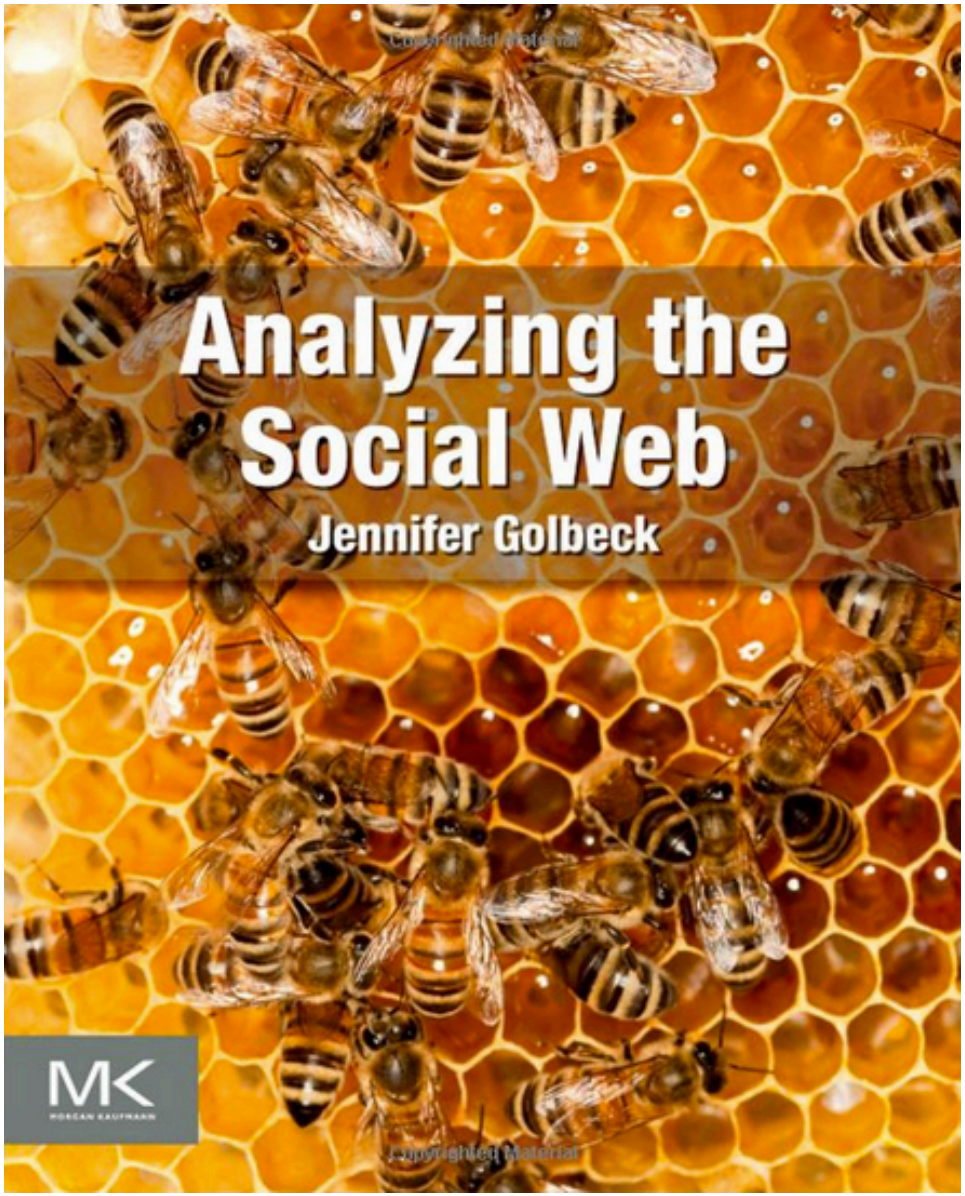


Source: http://www.amazon.com/Python-Finance-Analyze-Financial-Data/dp/1491945281

Business Insights with Social Analytics

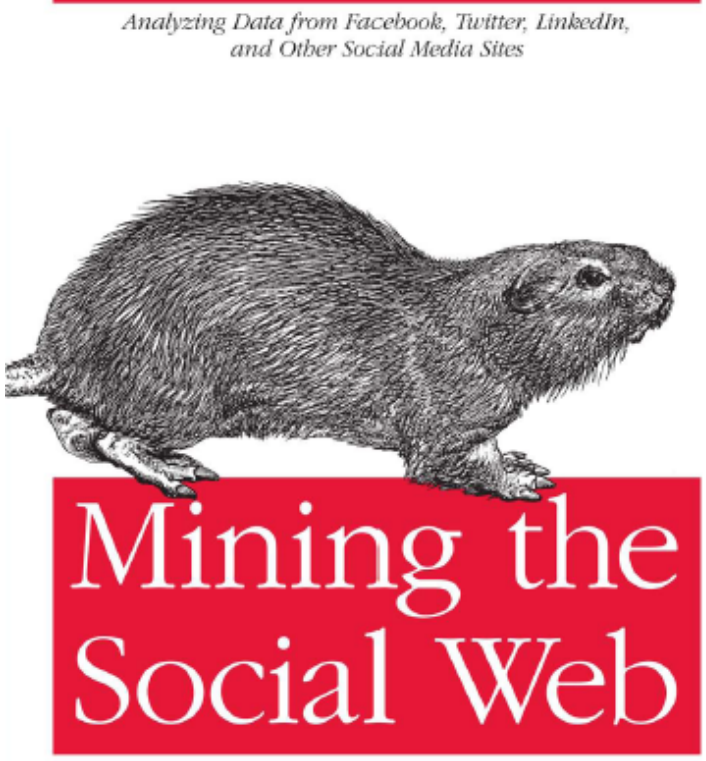
Analyzing the Social Web:
Social Network Analysis

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



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

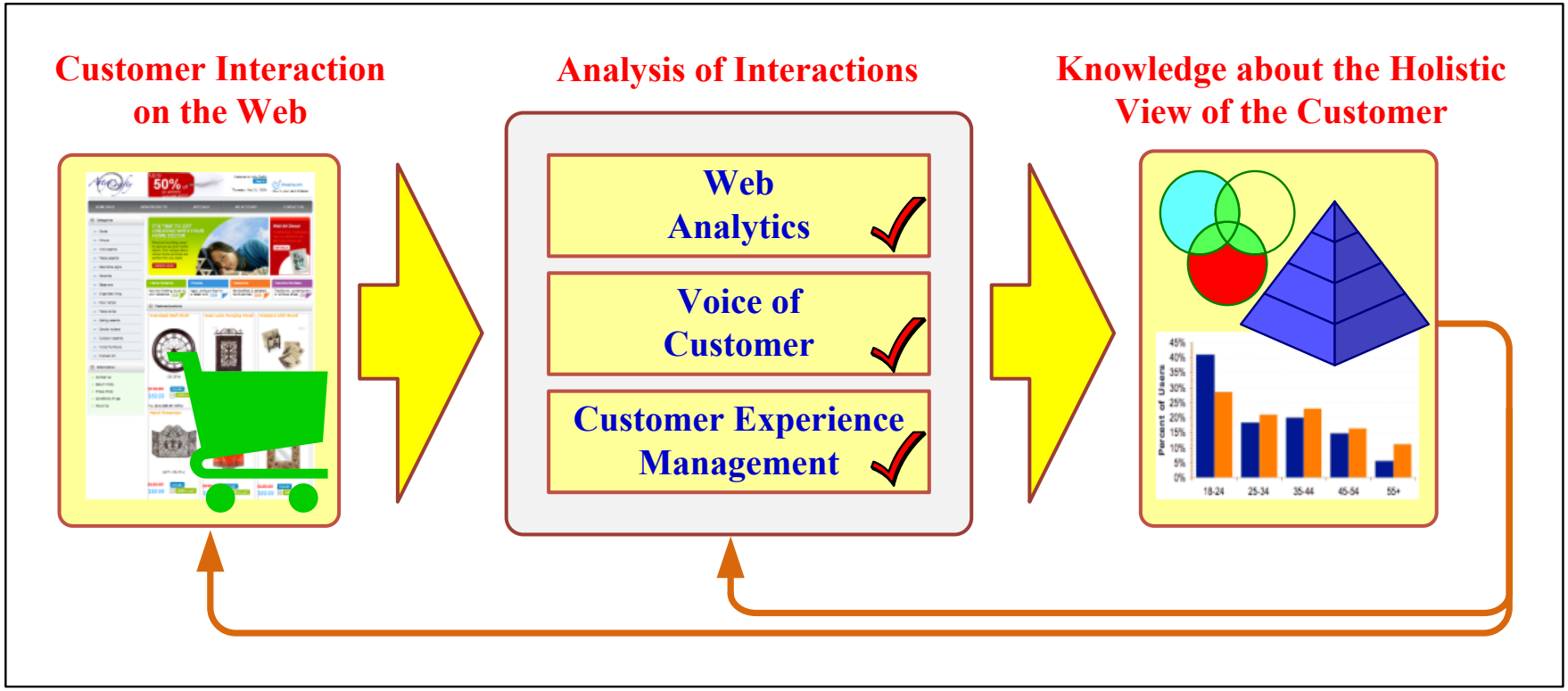
Mining the Social Web: Analyzing Data from Facebook, Twitter, LinkedIn, and Other Social Media Sites



Source: http://www.amazon.com/Mining-Social-Web-Analyzing-Facebook/dp/1449388345

Web Mining Success Stories

- Amazon.com, Ask.com, Scholastic.com, …
- Website Optimization Ecosystem



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

Business Intelligence Trends

1. Agile Information Management (IM)
2. Cloud Business Intelligence (BI)
3. Mobile Business Intelligence (BI)
4. Analytics
5. Big Data

Source: http://www.businessspectator.com.au/article/2013/1/22/technology/five-business-intelligence-trends-2013

Business Intelligence Trends: Computing and Service

- Cloud Computing and Service
- Mobile Computing and Service
- Social Computing and Service

Business Intelligence and Analytics

- Business Intelligence 2.0 (BI 2.0)
  - Web Intelligence
  - Web Analytics
  - Web 2.0
  - Social Networking and Microblogging sites
- Data Trends
  - Big Data
- Platform Technology Trends
  - Cloud computing platform

Source: Lim, E. P., Chen, H., & Chen, G. (2013). Business Intelligence and Analytics: Research Directions. *ACM Transactions on Management Information Systems (TMIS)*, *3*(4), 17

Business Intelligence and Analytics: Research Directions

1. Big Data Analytics
   - Data analytics using Hadoop / MapReduce framework
2. Text Analytics
   - From Information Extraction to Question Answering
   - From Sentiment Analysis to Opinion Mining
3. Network Analysis
   - Link mining
   - Community Detection
   - Social Recommendation

Source: Lim, E. P., Chen, H., & Chen, G. (2013). Business Intelligence and Analytics: Research Directions. *ACM Transactions on Management Information Systems (TMIS)*, *3*(4), 17

Source: McAfee, A., & Brynjolfsson, E. (2012). Big data: the management revolution.*Harvard business review*.

Data Scientist:

The Sexiest Job of the 21st Century

Meet the people who can coax treasure out of messy, unstructured data.

by Thomas H. Davenport and D.J. Patil

When Jonathan Goldman arrived for work in June 2006 at LinkedIn, the business networking site, the place still felt like a start-up. The company had just under 8 million accounts, and the number was growing quickly as existing members invited their friends and colleagues to join. But users weren't seeking out connections with the people who were already on the site at the rate executives had expected. Something was apparently missing in the social experience. As one LinkedIn manager put it, "It was like arriving at a conference reception and realizing you don't know anyone. So you just stand in the corner sipping your drink—and you probably leave early."

Source: Davenport, T. H., & Patil, D. J. (2012). Data Scientist. *Harvard business review*

AI Challenge Champion

SAS第七屆大數據資料科學家競賽

FinTech跨界整合X未來模型-資料時代領航者挑戰賽

FinTech跨界整合X未來模型-資料時代領航者挑戰賽

在這個巨量資料的時代，懂得巨量分析的專業人才「資料科學家」（Data Scientist）將成為未來炙手可熱的明日之星。 SAS 希望學生以創意無限及發掘新商機的角度出發，搭配巨量資料分析實例主題，鼓勵全國大學以分組專案及簡報競賽方式，分析高達上億筆的巨量資料，親身體驗巨量分析的神奇魔力!

早鳥報名・優惠方案

報名成功者，並於2018年3月9日前匯款完畢
即享有八折早鳥報名優惠!
（原報名費每隊1000元，早鳥優惠價每隊800元）

我要報名

http://saschampion.com.tw/

SAS第七屆大數據資料科學家競賽

FinTech跨界整合X未來模型-資料時代領航者挑戰賽

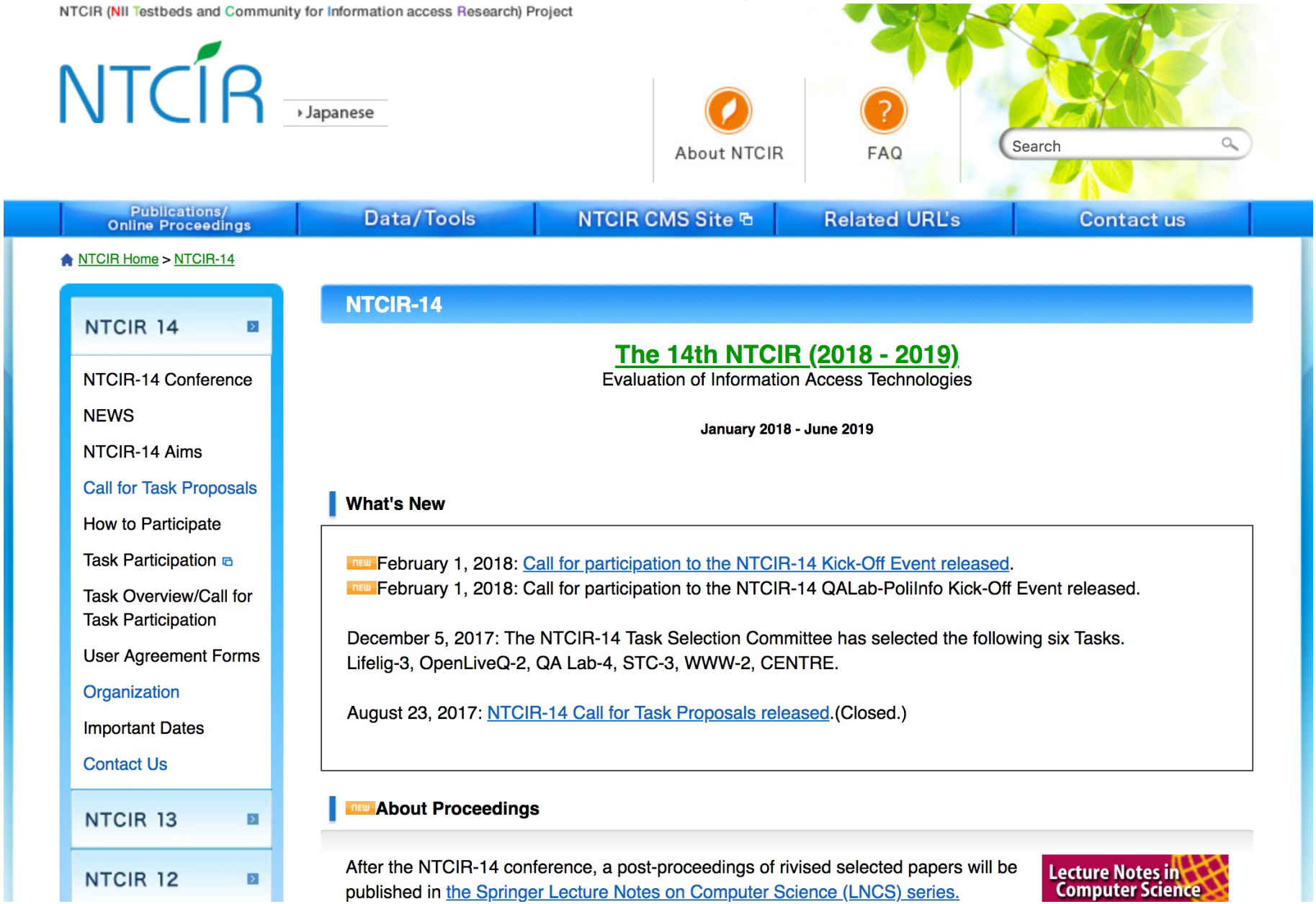
A. 賽制說明

本次報名組別將分為「模型挑戰組」與「商業洞察組」，同學可依照自身較擅長的優勢選擇其中一組別進行報名，共 150 組額滿為止，但任一組別報名超過 85 組即無法報名！

	模型挑戰組	商業洞察組
建議挑戰族群 【不強制任何科系需選特定組別】	對模型分析、演算法分析有熱忱及創意想法的學生族群。 ★建議的背景科系： 統計、資訊學科類及工程學科相關科系	對分析結合應用有熱忱、有商業敏感度的學生族群。 ★建議的背景科系： 統計、資訊學科類、商管學科相關科系
初賽評分標準	分別取初賽平均分數**前20名**組別晉級複賽	
複賽評分標準	分別取「探索式資料分析報告」評分**前7名**進入決賽	
決賽評分標準	複賽成績 25% 模型準度 50% 商業模式 25%	複賽成績 25% 模型準度 25% 商業模式 50%

http://saschampion.com.tw/detail.php

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

Short Text Conversation Task (STC-3)

Chinese Emotional Conversation Generation (CECG) Subtask

- Home
- Task Definition
- Dataset Description
- Evaluation Metric
- Time Schedule
- Copy Rights & Contacts

Links

- NTCIR-14
- NTCIR-14 STC-3
- NLPCC 2017

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.

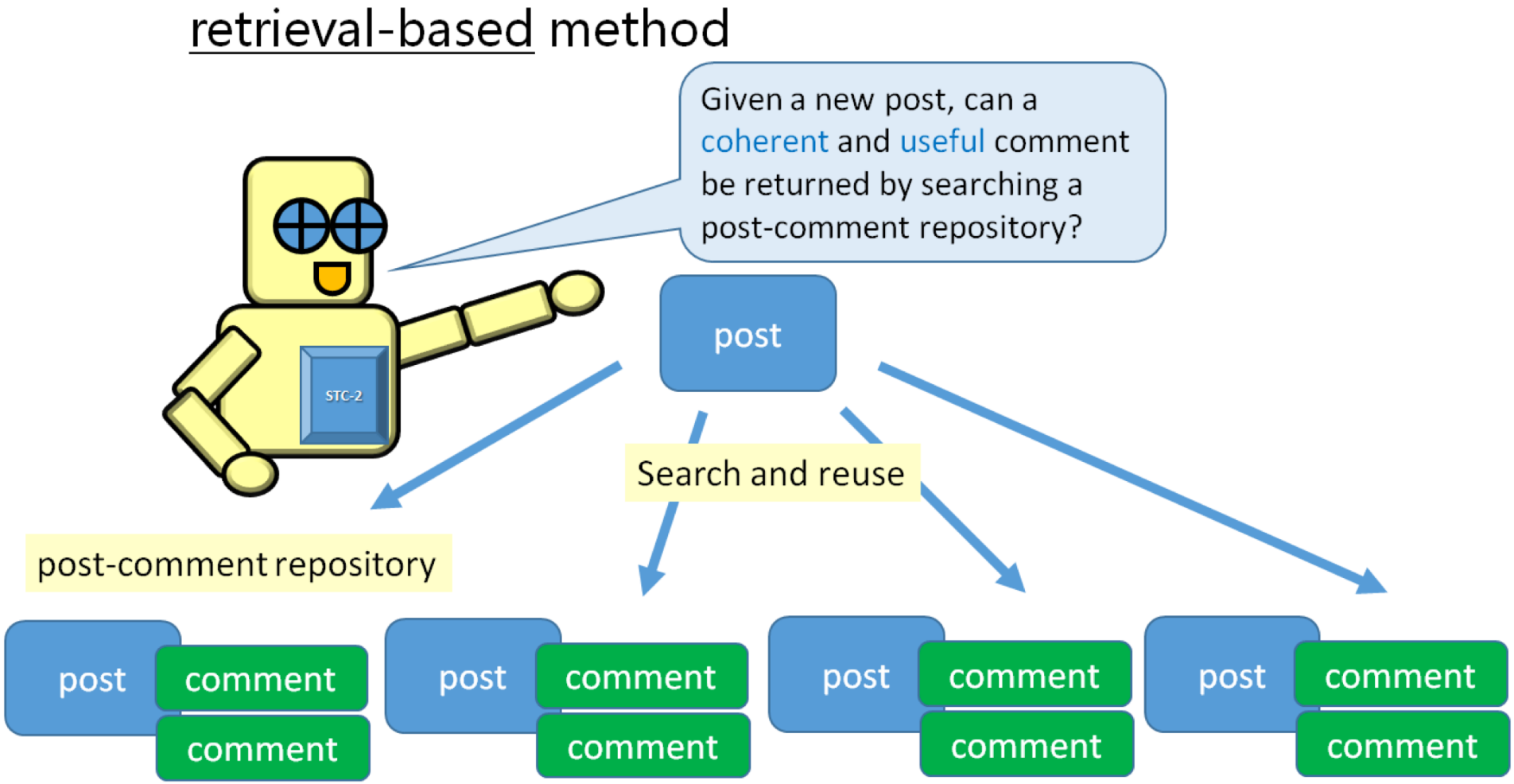
In this challenge, participants are expected to generate Chinese responses that are not only appropriate in content but also adequate in emotion, which is quite important for building an empathic chatting machine. For instance, if user says “My cat died yesterday”, the most appropriate response may be “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

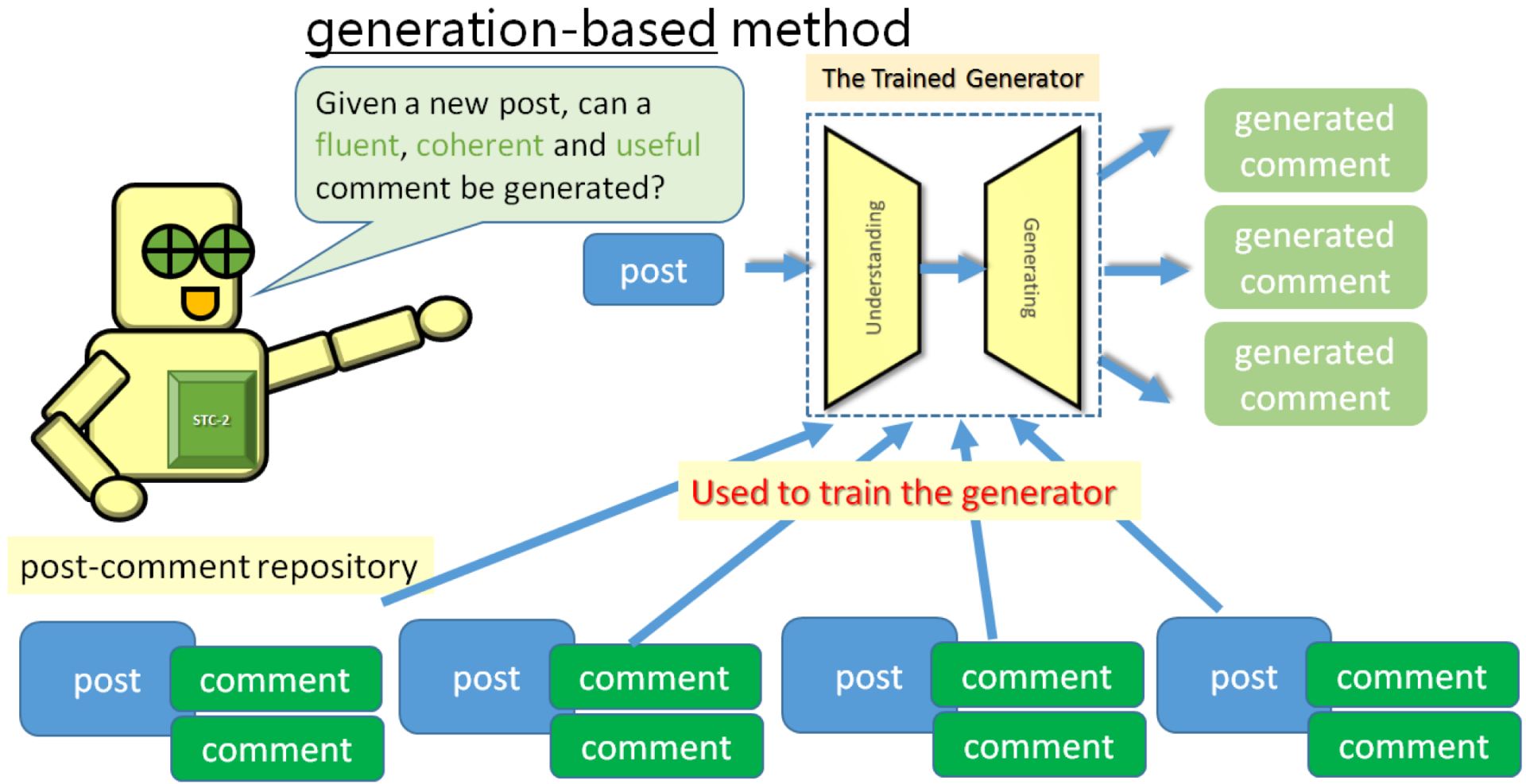
http://www.aihuang.org/p/challenge.html

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

科技大擂台 與AI對話 (正式賽)

科技大擂台 與AI對話(正式賽)

報名時間：2017-12-27～2018-02-28

人工智慧（AI）正加速改變全球產業、經濟與社會生活發展型態，亦成為各大產業的發展重點。各項AI技術研發項目中，尤以語音應用為最重要的技術，因為語音對話是人機互動最直覺、最人性化的方式，語意理解技術是AI智慧應用的核心。科技部舉辦台灣首屆「科技大擂台 與AI對話」，以獎勵賽的模式鼓勵創新者運用創意與技術來解決語音AI應用的挑戰。

一、競賽目的

1. 建置多情境的中文語音大數據，提升我國AI團隊技術。
2. 加速中文語音對話的核心技術開發。

二、賽程規劃

三、參賽方式與資格

1. 初賽：本賽事採團體報名， 團員人數以10名為限。團隊中至少有一位中華民國公民，且其他成員若非本國籍者需持有中華民國工作許可或我國學籍，非本國籍成員報名時須檢附有效中華民國工作證明或在學證明，上述參賽資格需於本賽事報名截止日前(含當日)取得。詳細內容請參考簡章內容。

https://fgc.stpi.narl.org.tw/activity/techai2018

Summary

- This course introduces the fundamental concepts and applications technology of big data mining.
- Topics include
  - Big Data Mining
  - Big Data, Artificial Intelligence and Deep Learning
  - Association Analysis
  - Classification and Prediction
  - Cluster Analysis
  - Data Mining Using SAS Enterprise Miner (SAS EM)
  - Case Study and Implementation of Big Data Mining

Contact Information

戴敏育 博士 (Min-Yuh Day, Ph.D.)

專任助理教授
淡江大學 資訊管理學系

電話：02-26215656 #2846
傳真：02-26209737
研究室：B929
地址： 25137 新北市淡水區英專路151號
Email： myday@mail.tku.edu.tw
網址：http://mail.tku.edu.tw/myday/