Tamkang University
淡江大學

Tamkang University

Big Data Mining
巨量資料探勘

AI人工智慧與大數據分析 (Artificial Intelligence and Big Data Analytics)

1082DM02
MI4 (M2244) (2744)
Tue 3, 4 (10:10-12:00) (B218)

Min-Yuh Day
戴敏育
Associate Professor
副教授

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

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

2020-03-10

課程大綱 (Syllabus)

週次 (Week)	日期 (Date)	內容(Subject/Topics)
1	2020/03/03	巨量資料探勘課程介紹 (Course Orientation for Big Data Mining)
2	2020/03/10	AI人工智慧與大數據分析 (Artificial Intelligence and Big Data Analytics)
3	2020/03/17	分群分析 (Cluster Analysis)
4	2020/03/24	個案分析與實作一 (SAS EM 分群分析)： Case Study 1 (Cluster Analysis - K-Means using SAS EM)
5	2020/03/31	關連分析 (Association Analysis)
6	2020/04/07	個案分析與實作二 (SAS EM 關連分析)： Case Study 2 (Association Analysis using SAS EM)
7	2020/04/14	分類與預測 (Classification and Prediction)
8	2020/04/21	期中報告 (Midterm Project Presentation)

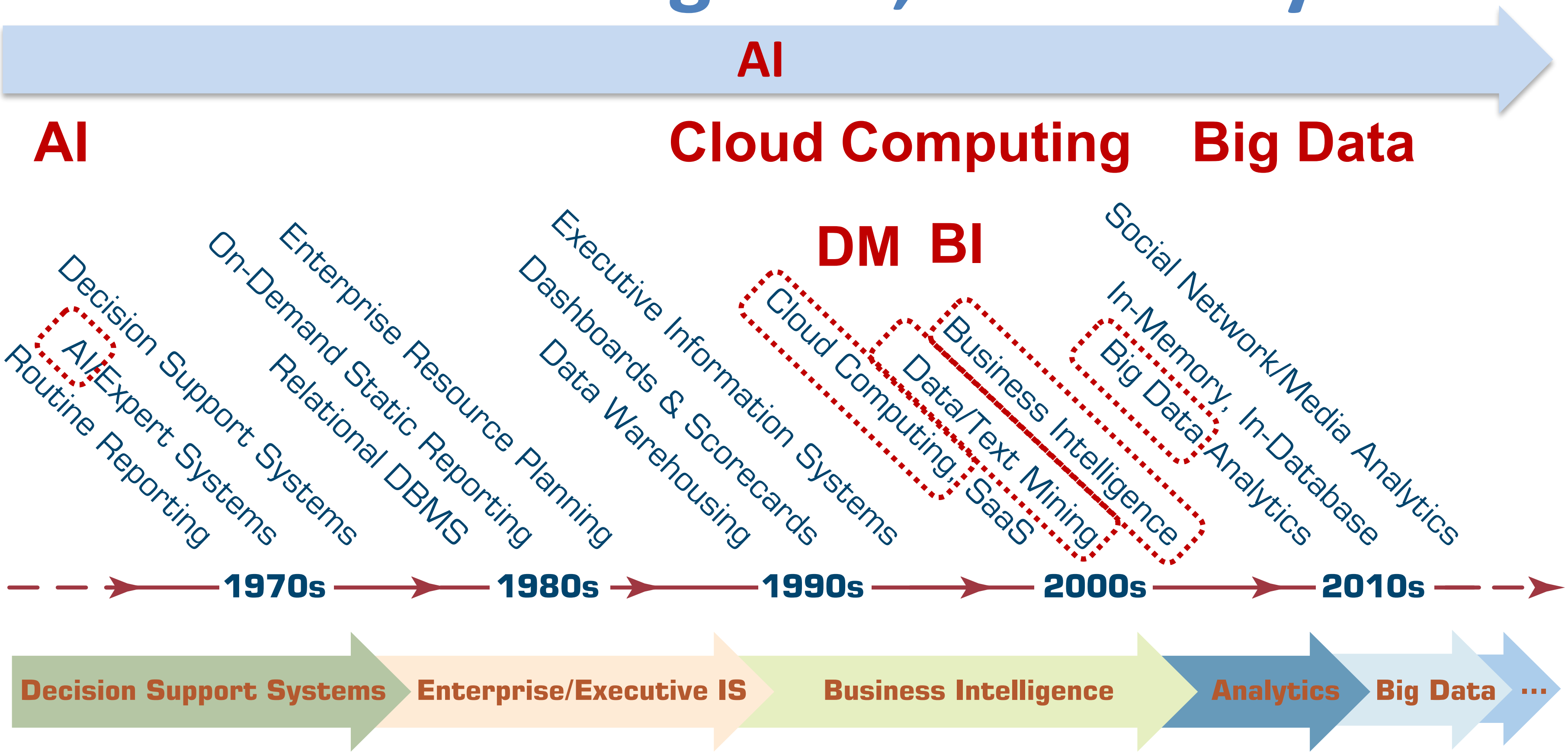
課程大綱 (Syllabus)

週次 (Week)	日期 (Date)	內容(Subject/Topics)
9	2020/04/28	期中考試週
10	2020/05/05	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)
11	2020/05/12	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)
12	2020/05/19	機器學習與深度學習 (Machine Learning and Deep Learning)
13	2020/05/26	期末報告 (Final Project Presentation)
14	2020/06/02	畢業考試週
15	2020/06/09	教師彈性補充教學

Outline

- AI
- Big Data Analytics

AI, Big Data, Cloud Computing
Evolution of Decision Support, Business Intelligence, and Analytics



AI

AI

Cloud Computing

Big Data

DM

BI

Routine Reporting

AI/Expert Systems

Decision Support Systems

Relational DBMS

On-Demand Static Reporting

Enterprise Resource Planning

Data Warehousing

Dashboards & Scorecards

Executive Information Systems

Cloud Computing, SaaS

Data/Text Mining

Business Intelligence

Big Data Analytics

In-Memory, In-Database

Social Network/Media Analytics

1970s → 1980s → 1990s → 2000s → 2010s

Decision Support Systems → Enterprise/Executive IS → Business Intelligence → Analytics → Big Data → …

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

AI

Definition of Artificial Intelligence (A.I.)

Artificial Intelligence

"... the **science** and

engineering

of

making

intelligent machines"

(John McCarthy, 1955)

Source: https://digitalintelligencetoday.com/artificial-intelligence-defined-useful-list-of-popular-definitions-from-business-and-science/

Artificial Intelligence

"… technology that thinks and acts like humans"

Source: https://digitalintelligencetoday.com/artificial-intelligence-defined-useful-list-of-popular-definitions-from-business-and-science/

Artificial Intelligence

"… intelligence exhibited by machines or software"

Source: https://digitalintelligencetoday.com/artificial-intelligence-defined-useful-list-of-popular-definitions-from-business-and-science/

4 Approaches of AI

Thinking Humanly	Thinking Rationally
Acting Humanly	Acting Rationally

Source: Stuart Russell and Peter Norvig (2016) , Artificial Intelligence: A Modern Approach, 3rd Edition, Pearson International

4 Approaches of AI

2. Thinking Humanly: The Cognitive Modeling Approach	3. Thinking Rationally: The “Laws of Thought” Approach
1. Acting Humanly: The Turing Test Approach (1950)	4. Acting Rationally: The Rational Agent Approach

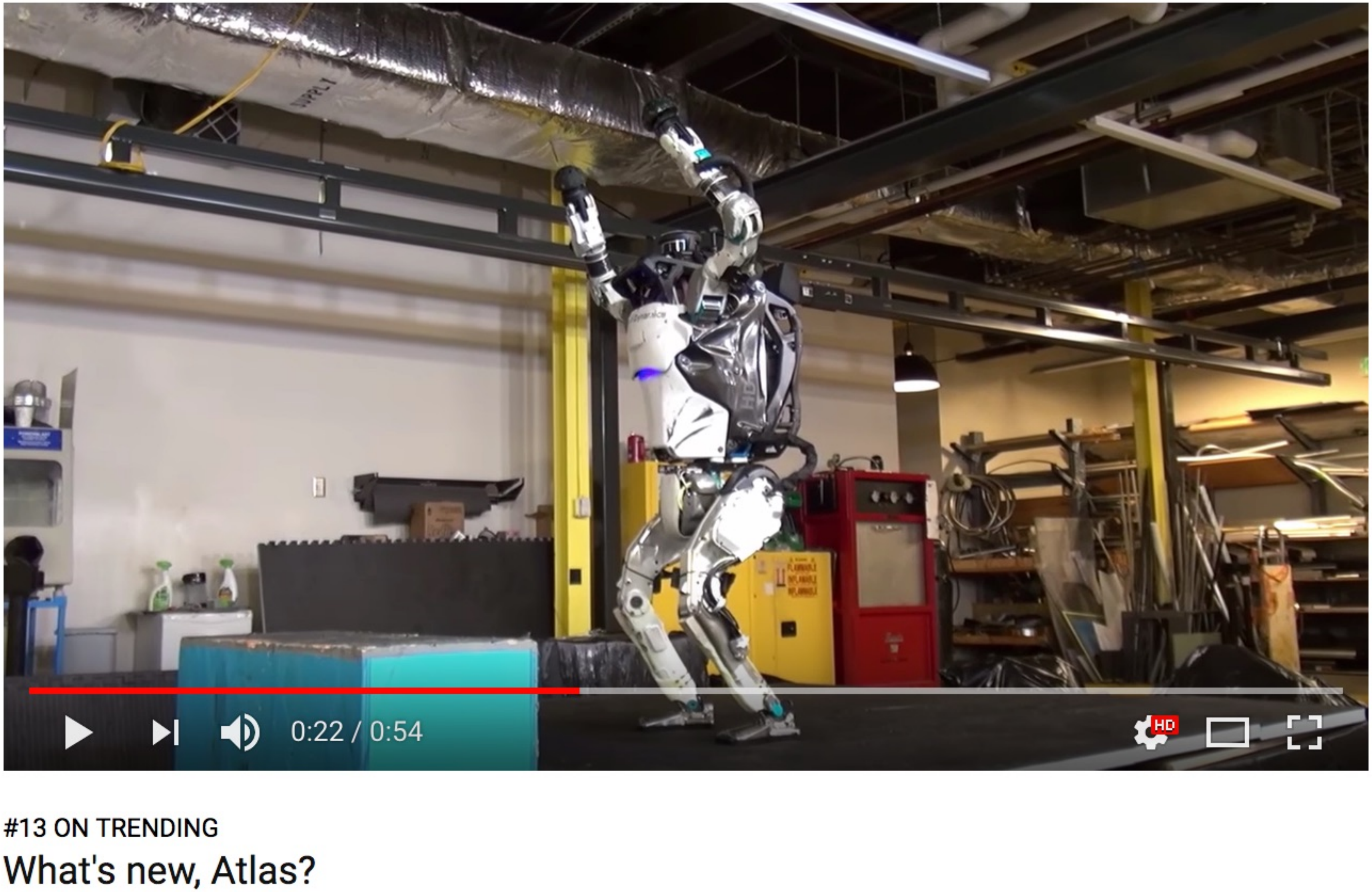
Source: Stuart Russell and Peter Norvig (2016) , Artificial Intelligence: A Modern Approach, 3rd Edition, Pearson International

AI Acting Humanly: The Turing Test Approach (Alan Turing, 1950)

- Natural Language Processing (NLP)
- Knowledge Representation
- Automated Reasoning
- Machine Learning (ML)
- Computer Vision
- Robotics

Source: Stuart Russell and Peter Norvig (2016) , Artificial Intelligence: A Modern Approach, 3rd Edition, Pearson International

Boston Dynamics: Atlas

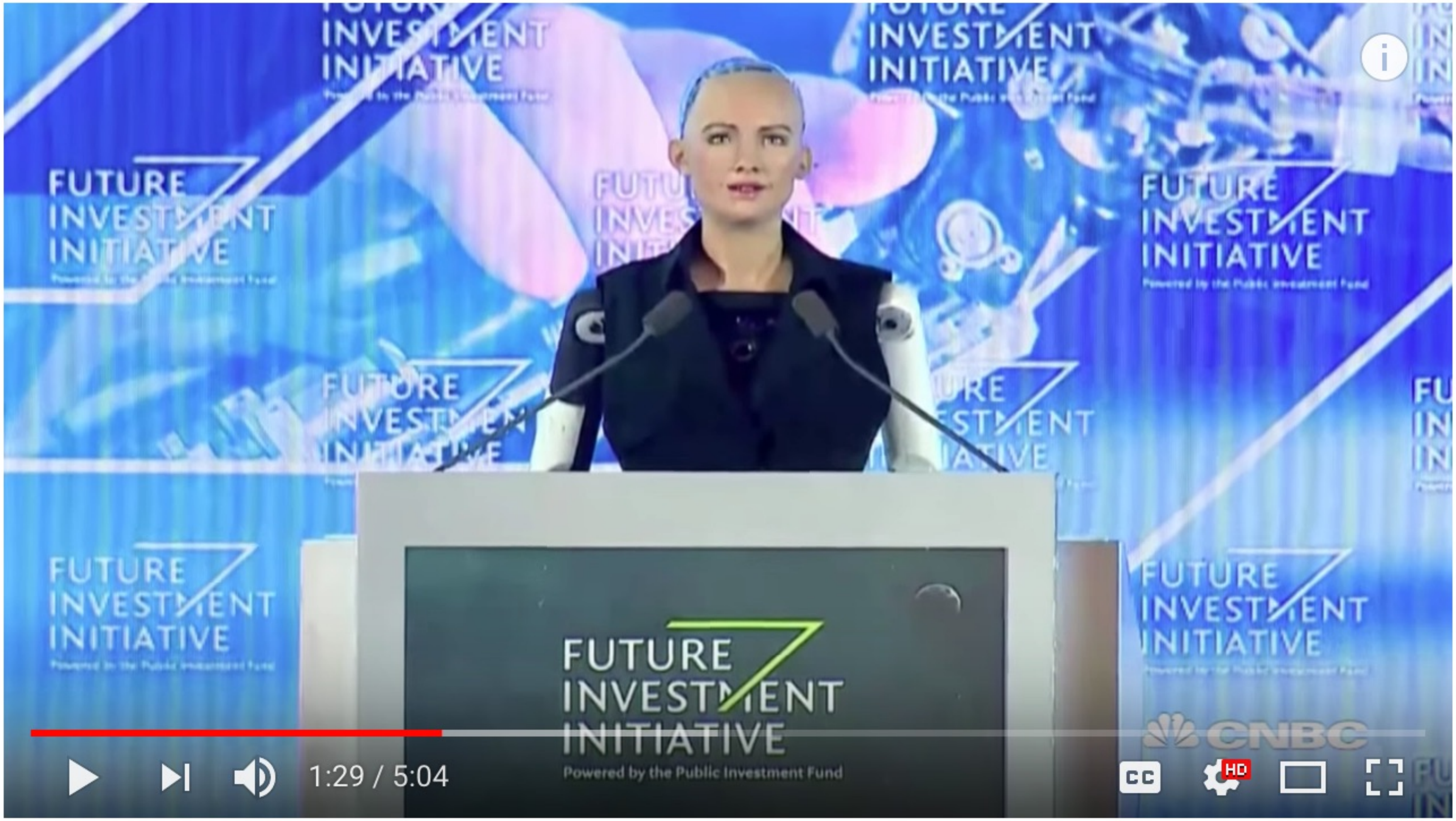


#13 ON TRENDING

What's new, Atlas?

https://www.youtube.com/watch?v=fRj34o4hN4I

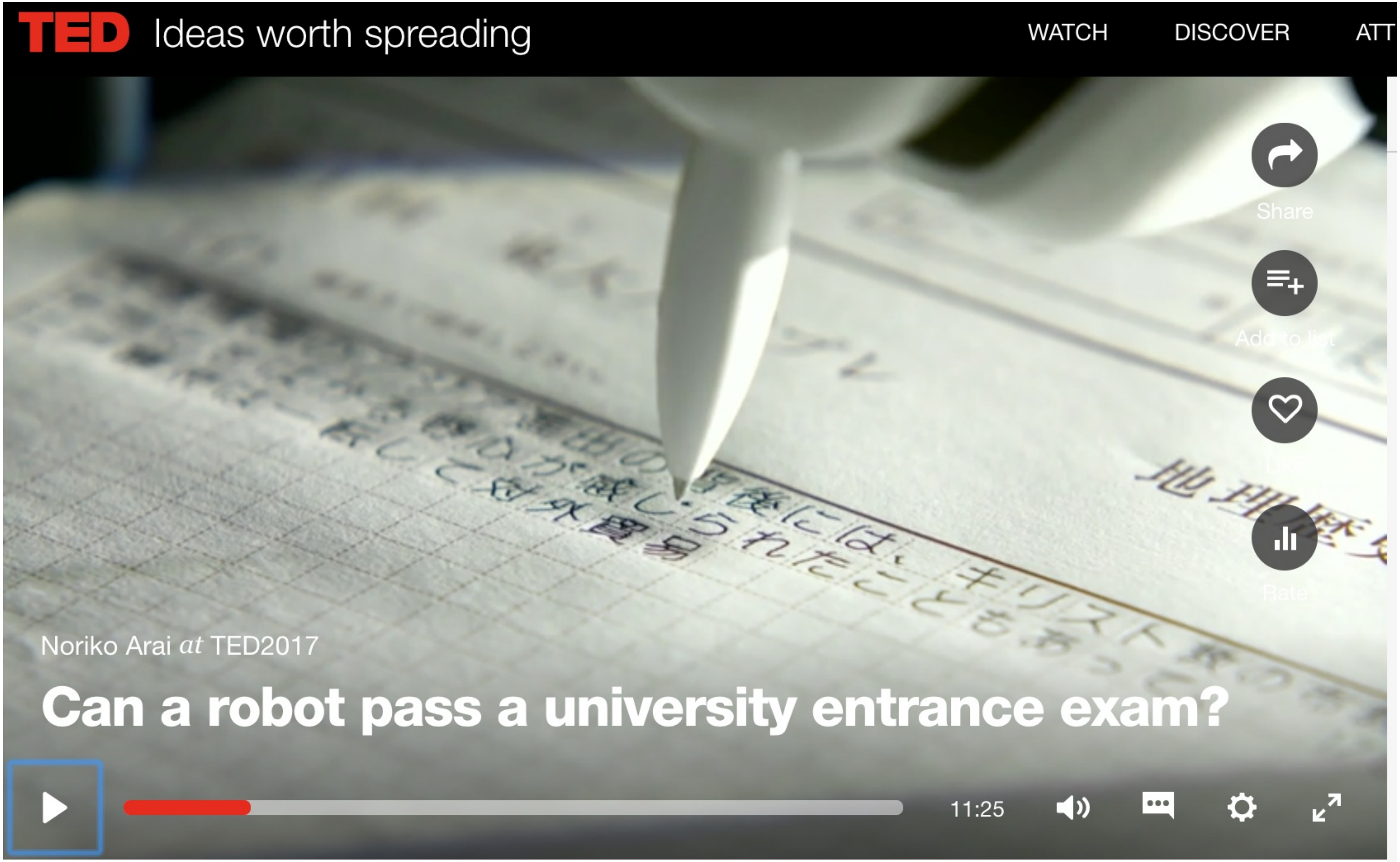
Humanoid Robot: Sophia


FUTURE INVESTMENT INITIATIVE
Powered by the Public Investment Fund
CNBC
1:29 / 5:04
CC
HD

https://www.youtube.com/watch?v=S5t6K9iwcdw

Can a robot pass a university entrance exam?

Noriko Arai at TED2017

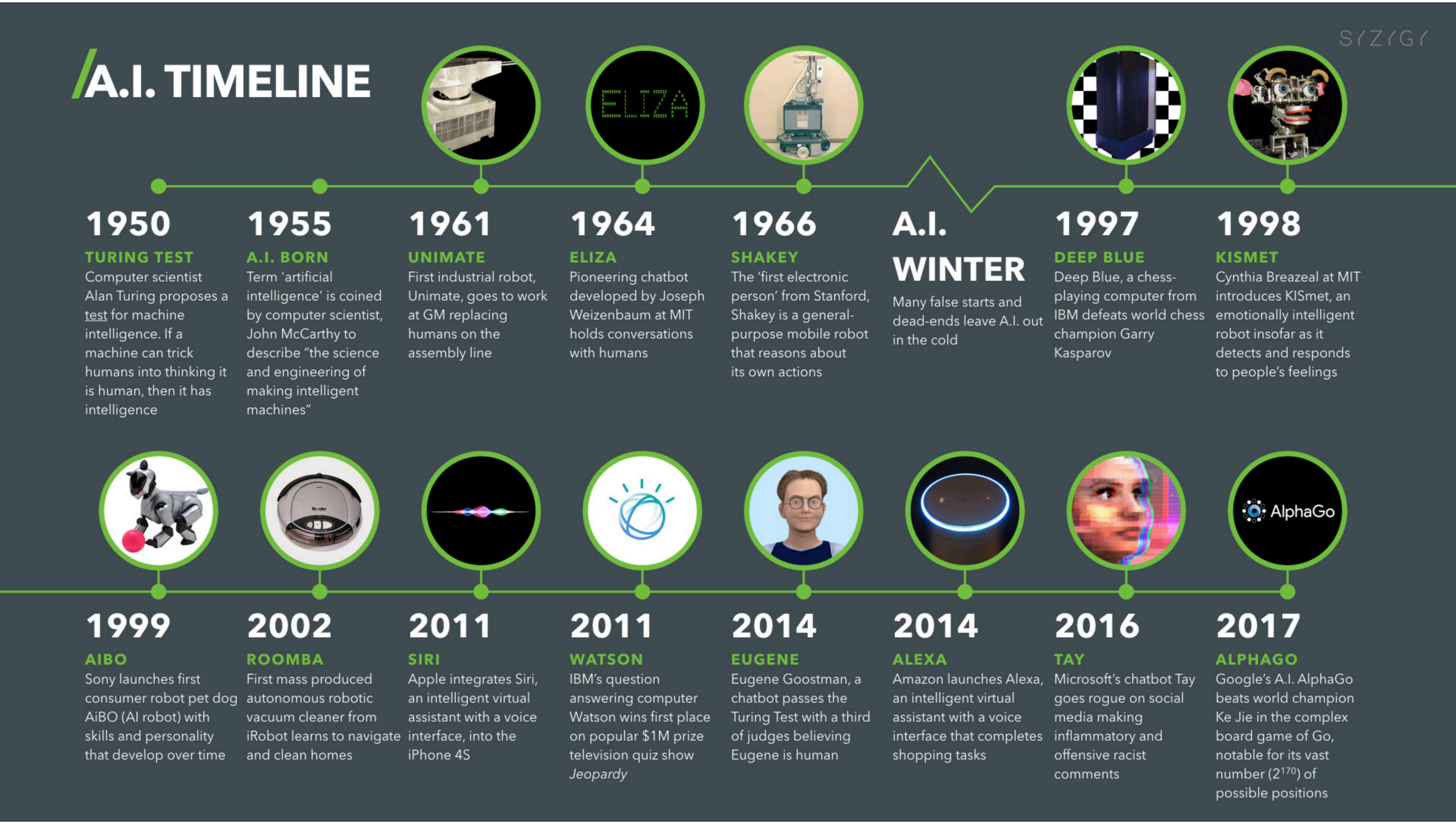


Noriko Arai *at* TED2017

Can a robot pass a university entrance exam?

https://www.ted.com/talks/noriko_arai_can_a_robot_pass_a_university_entrance_exam
https://www.youtube.com/watch?v=XQZjkPyJ8KU

Artificial Intelligence (A.I.) Timeline



A.I. TIMELINE

1950
TURING TEST
Computer scientist Alan Turing proposes a test for machine intelligence. If a machine can trick humans into thinking it is human, then it has intelligence

1955
A.I. BORN
Term 'artificial intelligence' is coined by computer scientist, John McCarthy to describe "the science and engineering of making intelligent machines"

1961
UNIMATE
First industrial robot, Unimate, goes to work at GM replacing humans on the assembly line

1964
ELIZA
Pioneering chatbot developed by Joseph Weizenbaum at MIT holds conversations with humans

1966
SHAKEY
The 'first electronic person' from Stanford, Shakey is a general-purpose mobile robot that reasons about its own actions

A.I. WINTER
Many false starts and dead-ends leave A.I. out in the cold

1997
DEEP BLUE
Deep Blue, a chess-playing computer from IBM defeats world chess champion Garry Kasparov

1998
KISMET
Cynthia Breazeal at MIT introduces KISmet, an emotionally intelligent robot insofar as it detects and responds to people's feelings

1999
AIBO
Sony launches first consumer robot pet dog AiBO (AI robot) with skills and personality that develop over time

2002
ROOMBA
First mass produced autonomous robotic vacuum cleaner from iRobot learns to navigate and clean homes

2011
SIRI
Apple integrates Siri, an intelligent virtual assistant with a voice interface, into the iPhone 4S

2011
WATSON
IBM's question answering computer Watson wins first place on popular $1M prize television quiz show *Jeopardy*

2014
EUGENE
Eugene Goostman, a chatbot passes the Turing Test with a third of judges believing Eugene is human

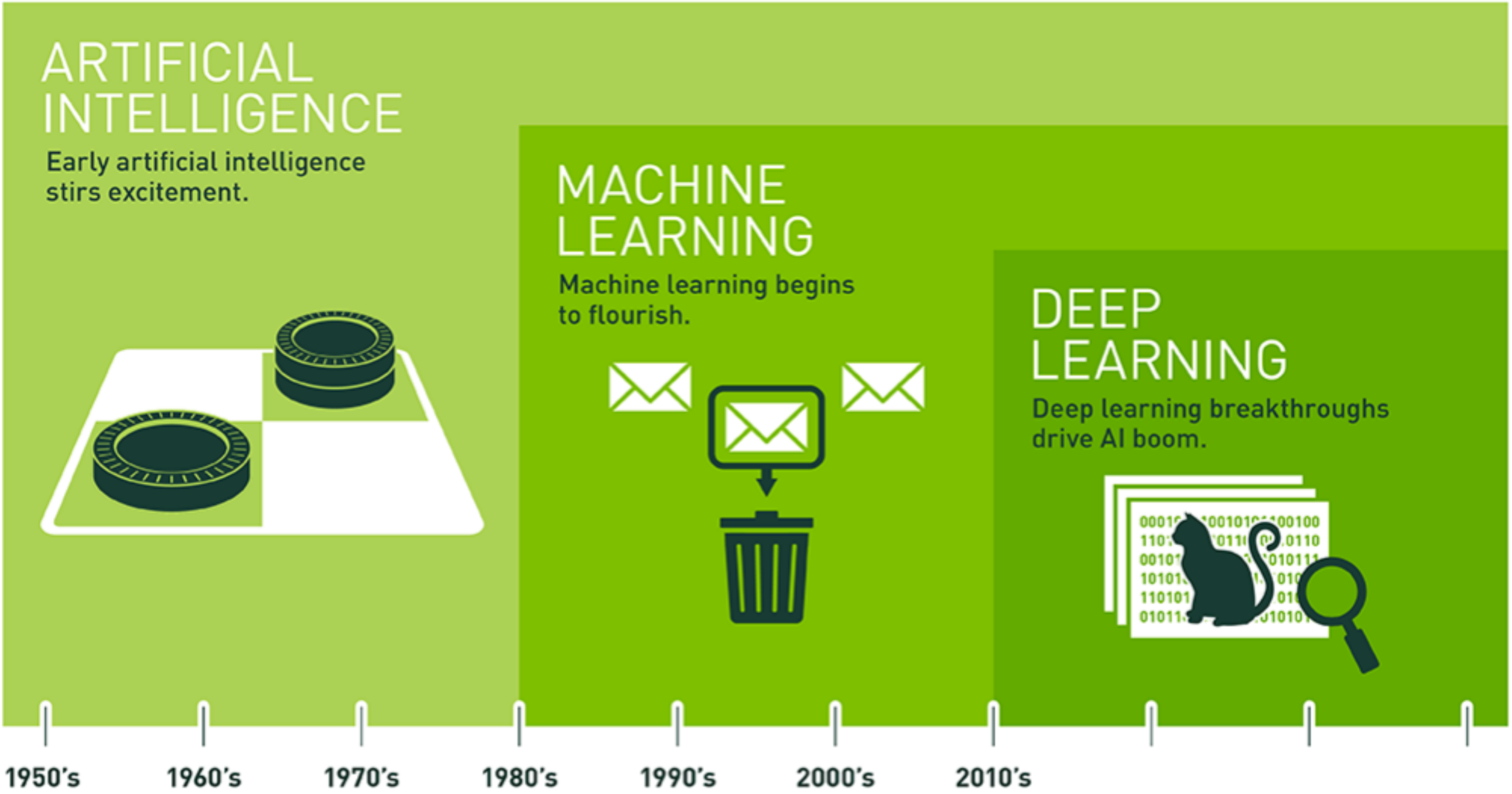
2014
ALEXA
Amazon launches Alexa, an intelligent virtual assistant with a voice interface that completes shopping tasks

2016
TAY
Microsoft's chatbot Tay goes rogue on social media making inflammatory and offensive racist comments

2017
ALPHAGO
Google's A.I. AlphaGo beats world champion Ke Jie in the complex board game of Go, notable for its vast number (2^{170}) of possible positions

Source: https://digitalintelligencetoday.com/artificial-intelligence-timeline-infographic-from-eliza-to-tay-and-beyond/

Artificial Intelligence
Machine Learning & Deep Learning



ARTIFICIAL INTELLIGENCE

Early artificial intelligence stirs excitement.

MACHINE LEARNING

Machine learning begins to flourish.

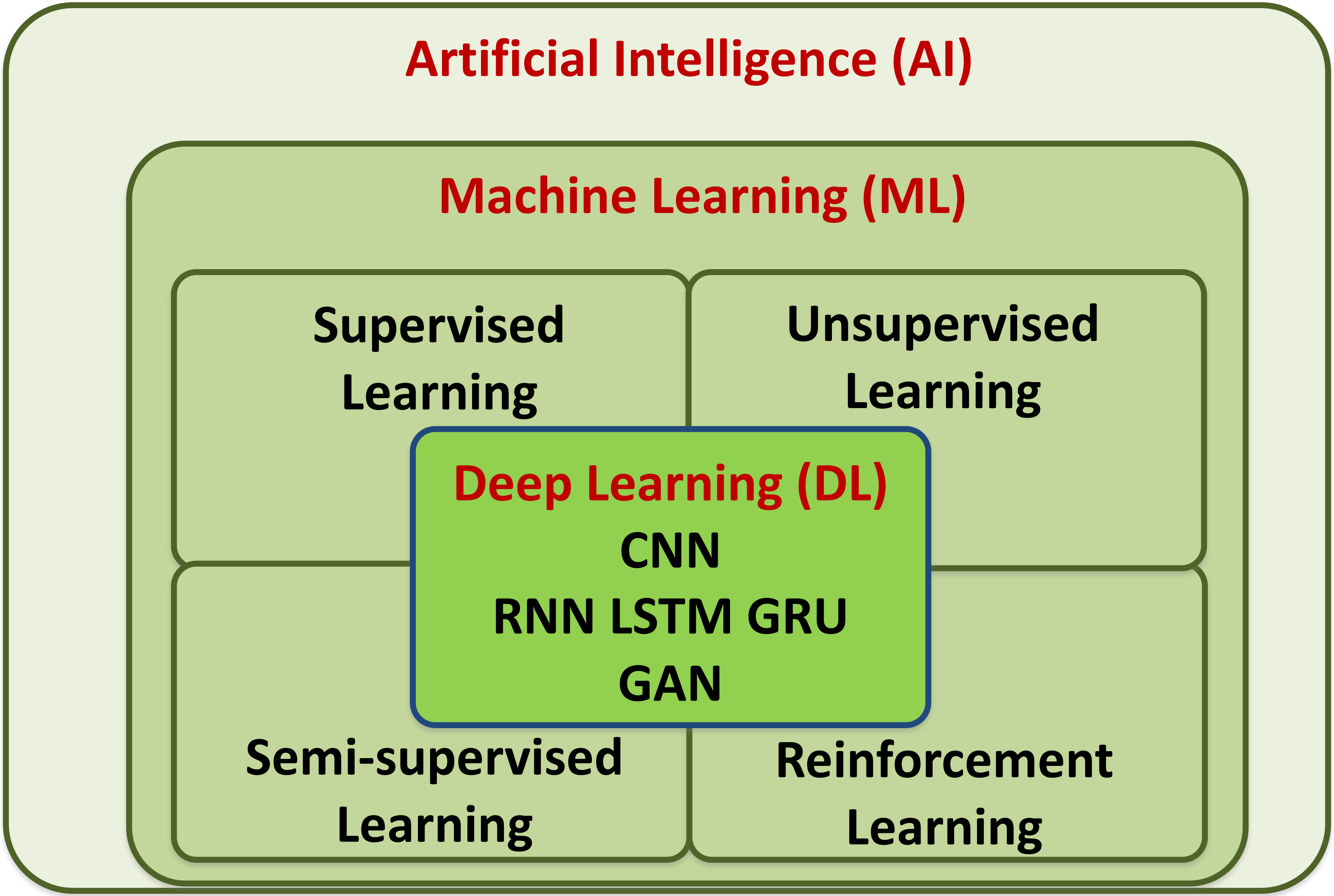
DEEP LEARNING

Deep learning breakthroughs drive AI boom.

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

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

AI, ML, DL



Artificial Intelligence (AI)

Machine Learning (ML)

Supervised Learning

Unsupervised Learning

Deep Learning (DL)

CNN

RNN LSTM GRU

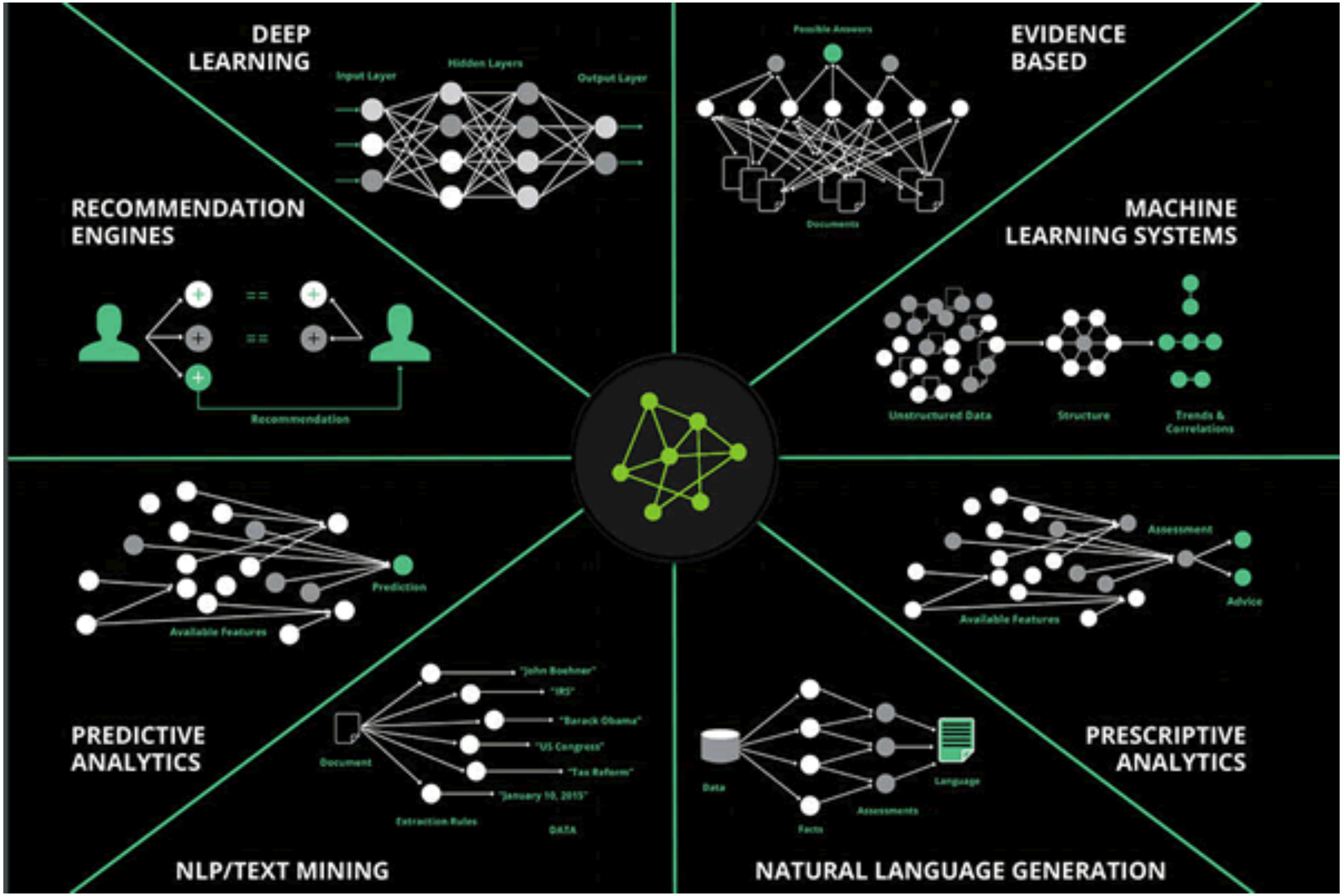
GAN

Semi-supervised Learning

Reinforcement Learning

Source: https://leonardoaraujosantos.gitbooks.io/artificial-inteligence/content/deep_learning.html

Artificial Intelligence (AI) is many things

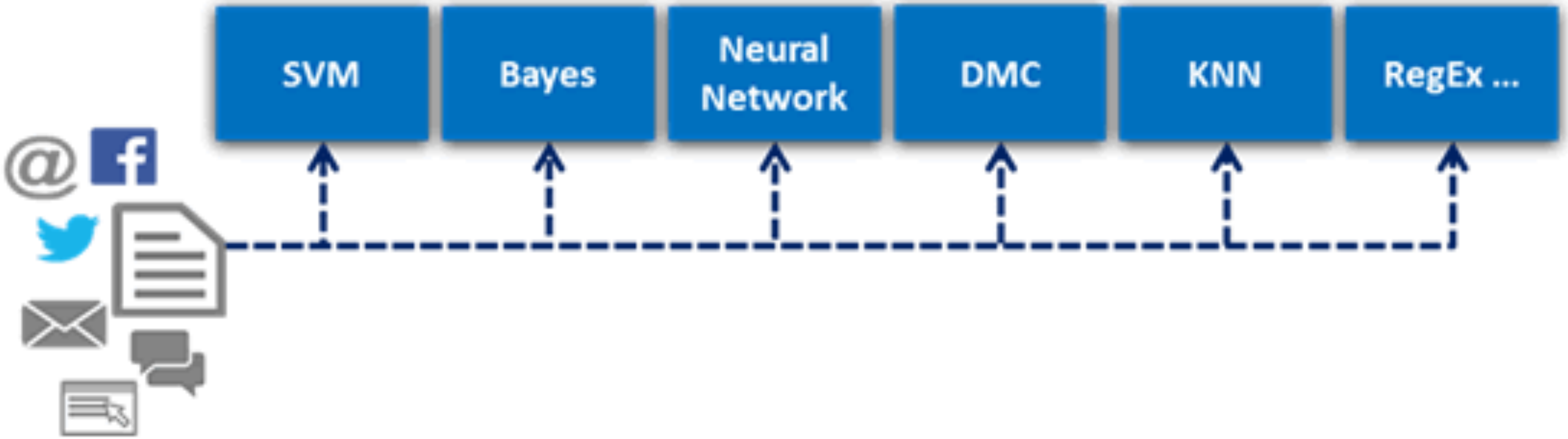


Ecosystem of AI

Source: https://www.i-scoop.eu/artificial-intelligence-cognitive-computing/

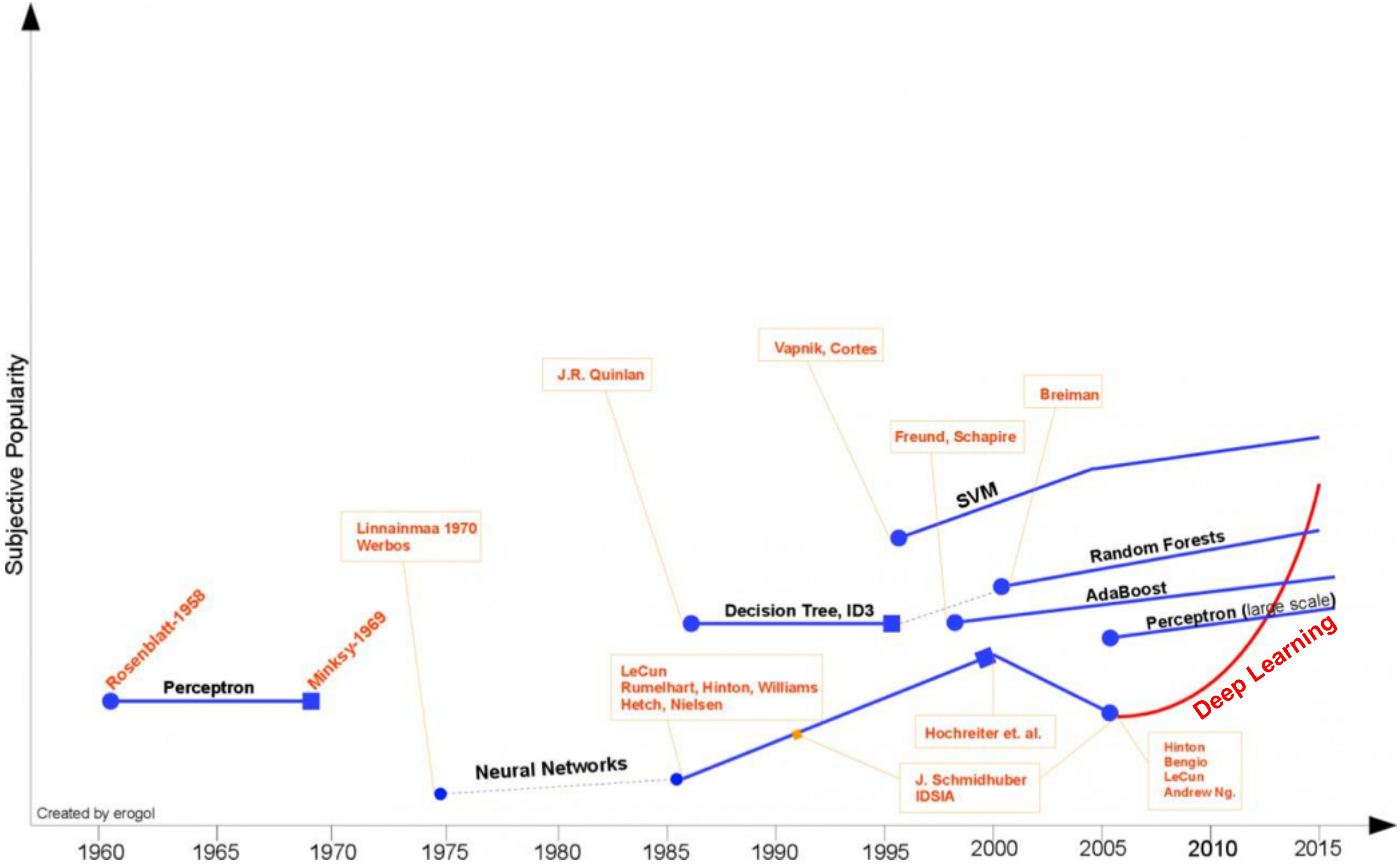
Artificial Intelligence (AI)

Intelligent Document Recognition algorithms



Source: https://www.i-scoop.eu/artificial-intelligence-cognitive-computing/

Deep Learning Evolution



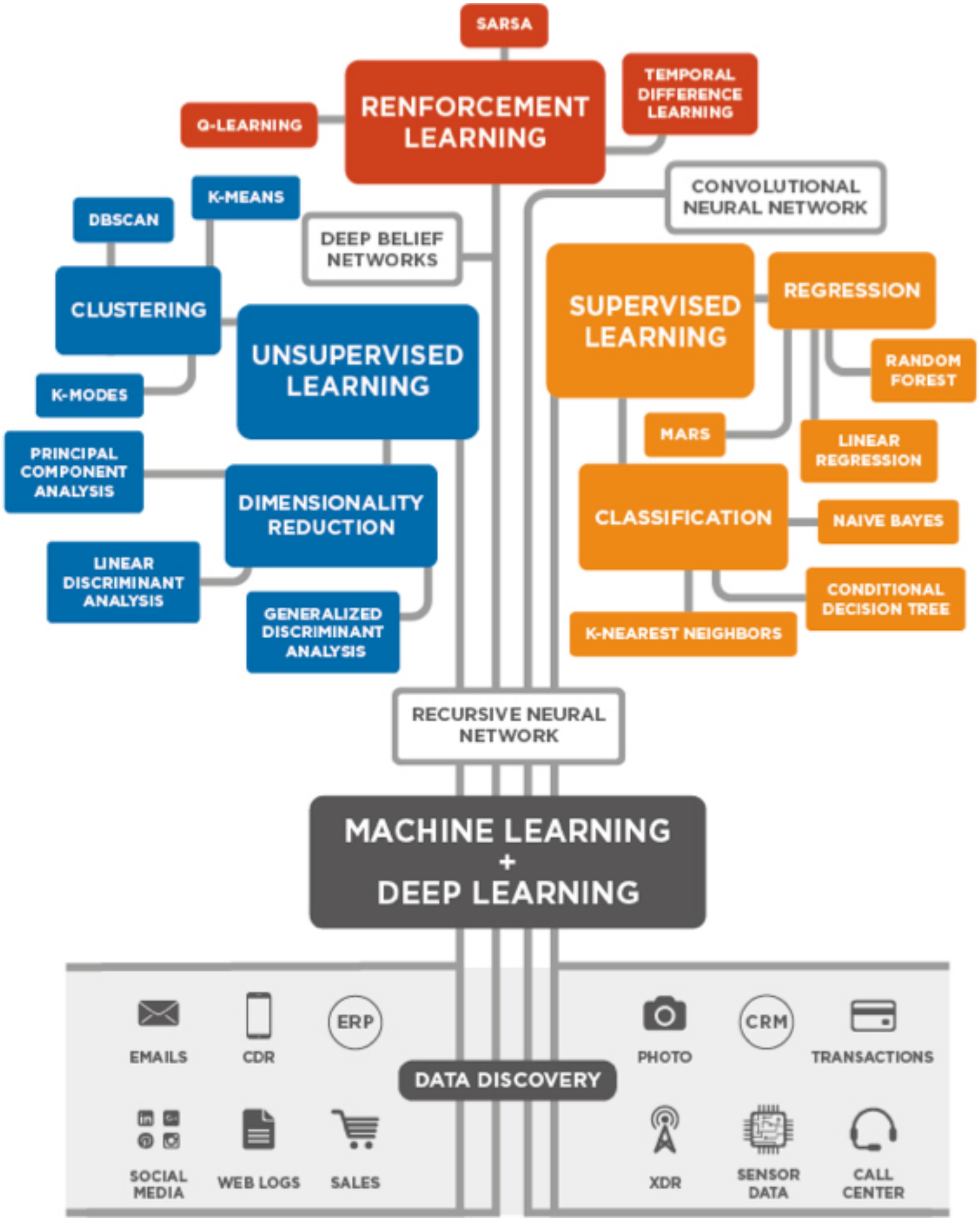
Source: http://www.erogol.com/brief-history-machine-learning/

Machine Learning Models

- Deep Learning
- Association rules
- Decision tree
- Clustering
- Bayesian
- Kernel
- Ensemble
- Dimensionality reduction
- Regression Analysis
- Instance based

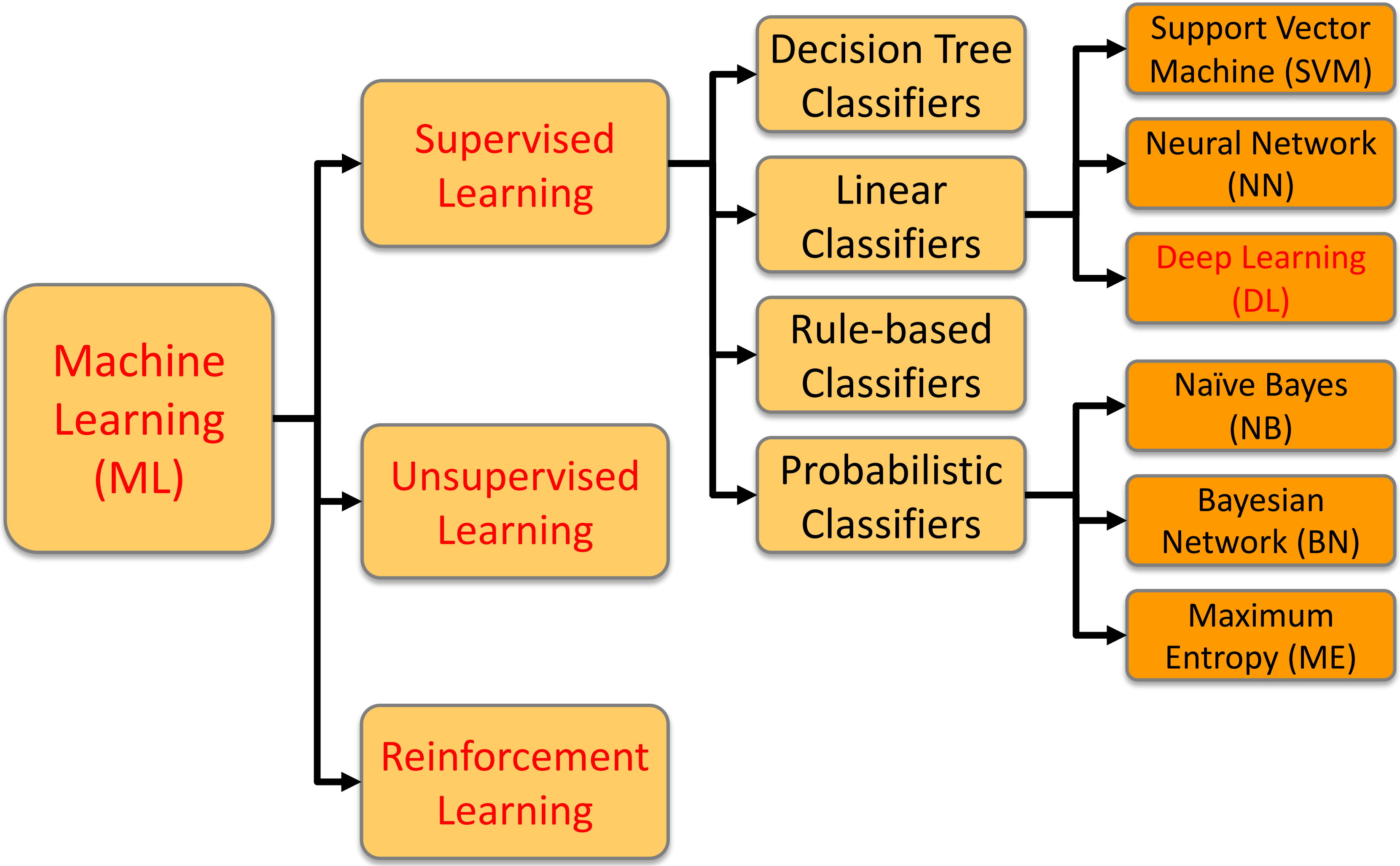
Source: Sunila Gollapudi (2016), Practical Machine Learning, Packt Publishing

3 Machine Learning Algorithms



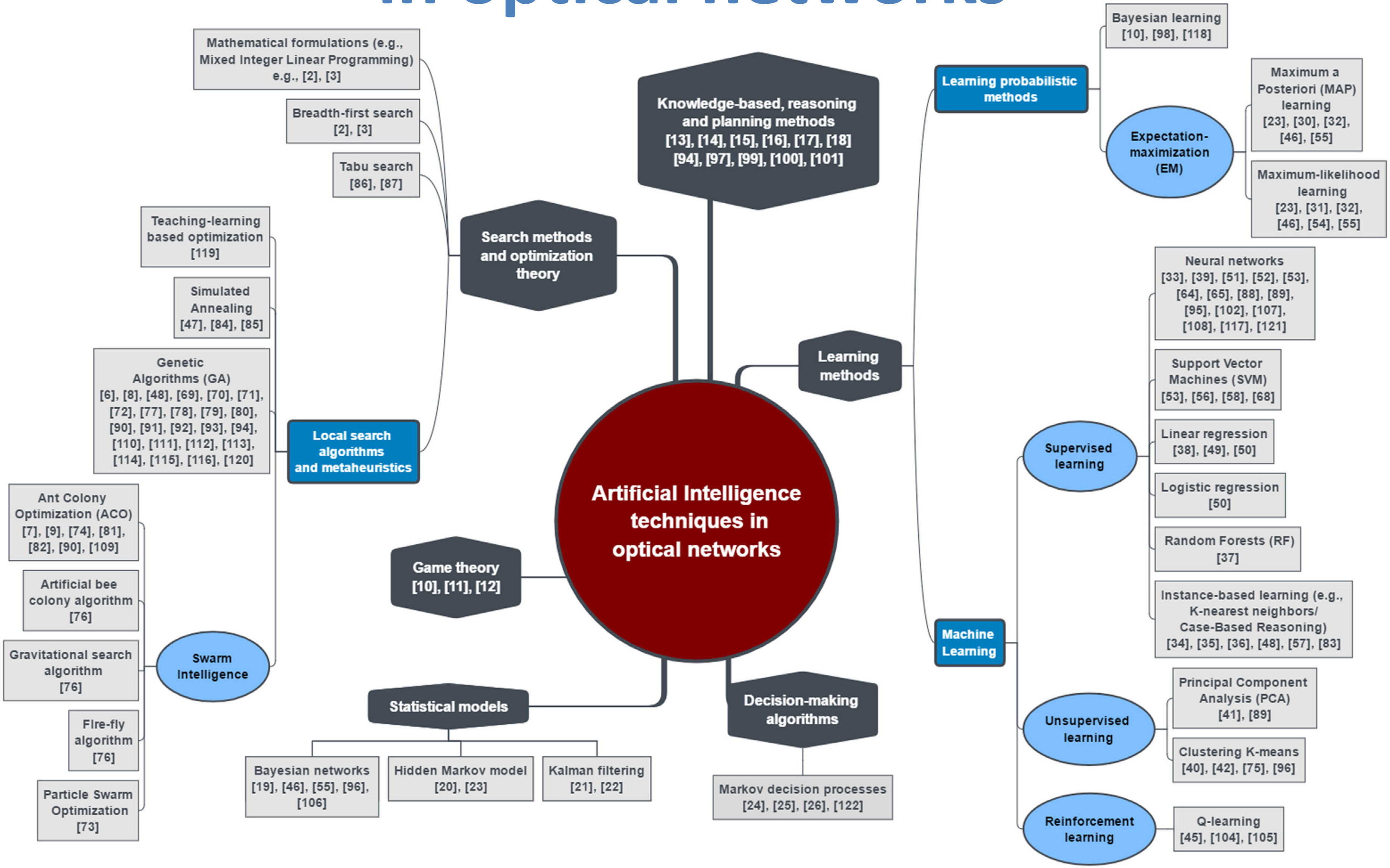
Source: Enrico Galimberti, http://blogs.teradata.com/data-points/tree-machine-learning-algorithms/

Machine Learning (ML) / Deep Learning (DL)



Source: Jesus Serrano-Guerrero, Jose A. Olivas, Francisco P. Romero, and Enrique Herrera-Viedma (2015), "Sentiment analysis: A review and comparative analysis of web services," Information Sciences, 311, pp. 18-38.

Artificial intelligence (AI) in optical networks

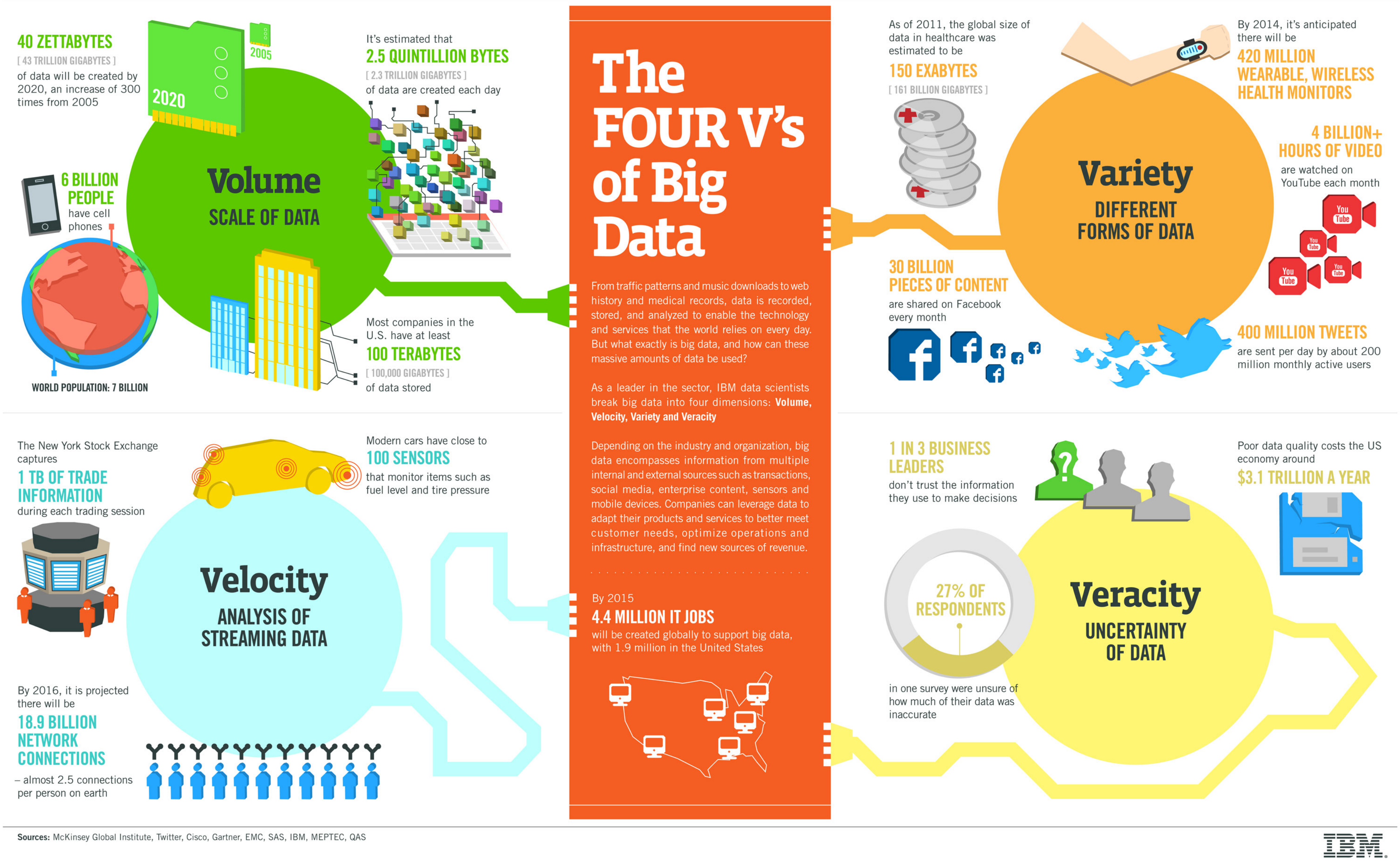


- Artificial Intelligence techniques in optical networks
  - Knowledge-based, reasoning and planning methods [13], [14], [15], [16], [17], [18] [94], [97], [99], [100], [101]
  - Search methods and optimization theory
    - Mathematical formulations (e.g., Mixed Integer Linear Programming) e.g., [2], [3]
    - Breadth-first search [2], [3]
    - Tabu search [86], [87]
    - Local search algorithms and metaheuristics
      - Teaching-learning based optimization [119]
      - Simulated Annealing [47], [84], [85]
      - Genetic Algorithms (GA) [6], [8], [48], [69], [70], [71], [72], [77], [78], [79], [80], [90], [91], [92], [93], [94], [110], [111], [112], [113], [114], [115], [116], [120]
      - Swarm Intelligence
        - Ant Colony Optimization (ACO) [7], [9], [74], [81], [82], [90], [109]
        - Artificial bee colony algorithm [76]
        - Gravitational search algorithm [76]
        - Fire-fly algorithm [76]
        - Particle Swarm Optimization [73]
  - Game theory [10], [11], [12]
  - Statistical models
    - Bayesian networks [19], [46], [55], [96], [106]
    - Hidden Markov model [20], [23]
    - Kalman filtering [21], [22]
  - Decision-making algorithms
    - Markov decision processes [24], [25], [26], [122]
  - Learning methods
    - Learning probabilistic methods
      - Bayesian learning [10], [98], [118]
      - Expectation-maximization (EM)
        - Maximum a Posteriori (MAP) learning [23], [30], [32], [46], [55]
        - Maximum-likelihood learning [23], [31], [32], [46], [54], [55]
    - Machine Learning
      - Supervised learning
        - Neural networks [33], [39], [51], [52], [53], [64], [65], [88], [89], [95], [102], [107], [108], [117], [121]
        - Support Vector Machines (SVM) [53], [56], [58], [68]
        - Linear regression [38], [49], [50]
        - Logistic regression [50]
        - Random Forests (RF) [37]
        - Instance-based learning (e.g., K-nearest neighbors/ Case-Based Reasoning) [34], [35], [36], [48], [57], [83]
      - Unsupervised learning
        - Principal Component Analysis (PCA) [41], [89]
        - Clustering K-means [40], [42], [75], [96]
      - Reinforcement learning
        - Q-learning [45], [104], [105]

Source: Javier Mata, Ignacio de Miguel, Ramón J. Durán, Noemí Merayo, Sandeep Kumar Singh, Admela Jukan, and Mohit Chamania (2018),

Big Data Analytics and Data Mining

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

Volume

SCALE OF DATA

40 ZETTABYTES

[43 TRILLION GIGABYTES]

of data will be created by 2020, an increase of 300 times from 2005

It's estimated that

2.5 QUINTILLION BYTES

[2.3 TRILLION GIGABYTES]

of data are created each day

6 BILLION PEOPLE

have cell phones

WORLD POPULATION: 7 BILLION

Most companies in the U.S. have at least

100 TERABYTES

[100,000 GIGABYTES]

of data stored

Velocity

ANALYSIS OF STREAMING DATA

The New York Stock Exchange captures

1 TB OF TRADE INFORMATION

during each trading session

Modern cars have close to

100 SENSORS

that monitor items such as fuel level and tire pressure

By 2016, it is projected there will be

18.9 BILLION NETWORK CONNECTIONS

– almost 2.5 connections per person on earth

Variety

DIFFERENT FORMS OF DATA

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

150 EXABYTES

[161 BILLION GIGABYTES]

By 2014, it's anticipated there will be

420 MILLION WEARABLE, WIRELESS HEALTH MONITORS

4 BILLION+ HOURS OF VIDEO

are watched on YouTube each month

30 BILLION PIECES OF CONTENT

are shared on Facebook every month

400 MILLION TWEETS

are sent per day by about 200 million monthly active users

Veracity

UNCERTAINTY OF DATA

1 IN 3 BUSINESS LEADERS

don't trust the information they use to make decisions

Poor data quality costs the US economy around

$3.1 TRILLION A YEAR

27% OF RESPONDENTS

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

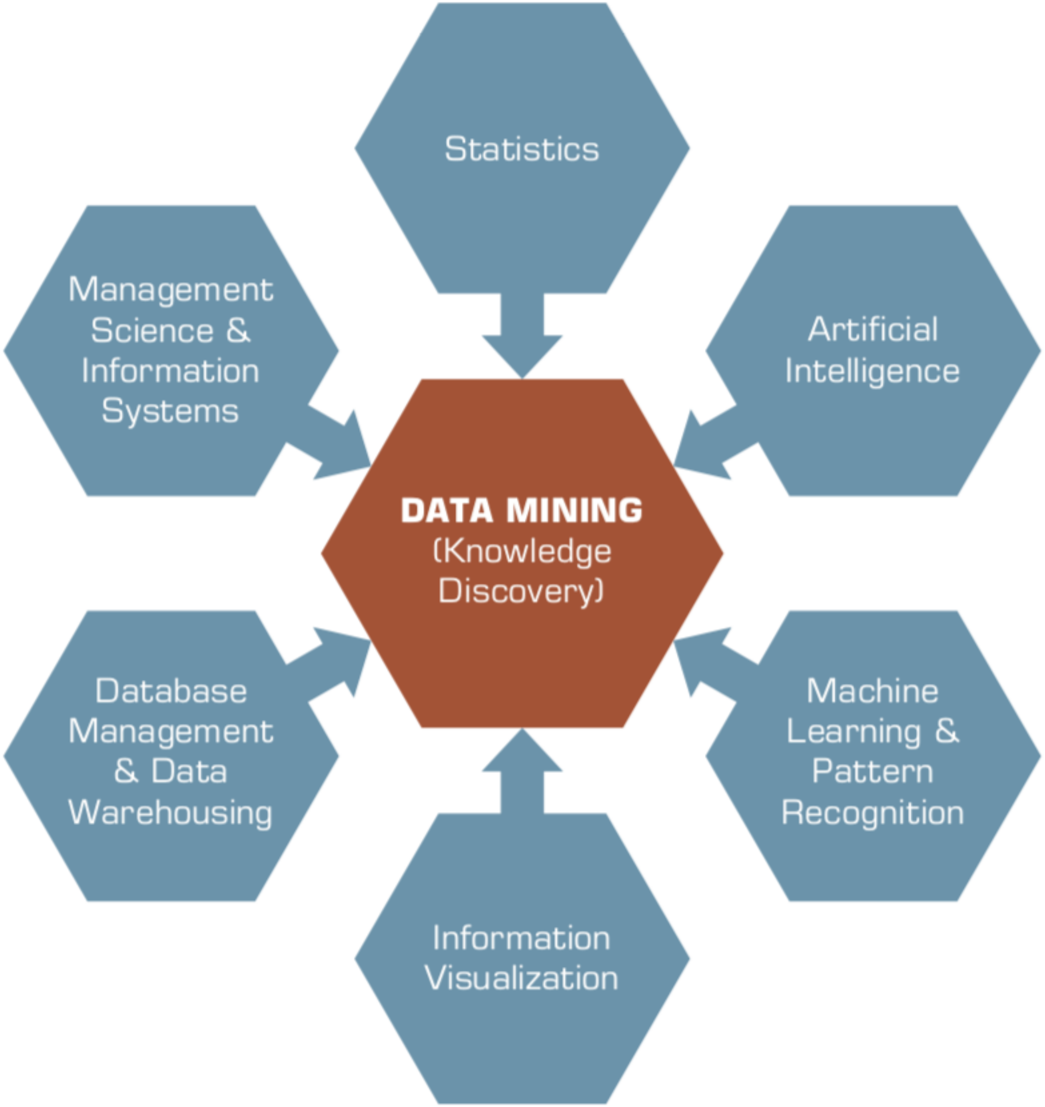
Sources: McKinsey Global Institute, Twitter, Cisco, Gartner, EMC, SAS, IBM, MEPTEC, QAS

IBM

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

Value

Data Mining
Is a Blend of Multiple Disciplines

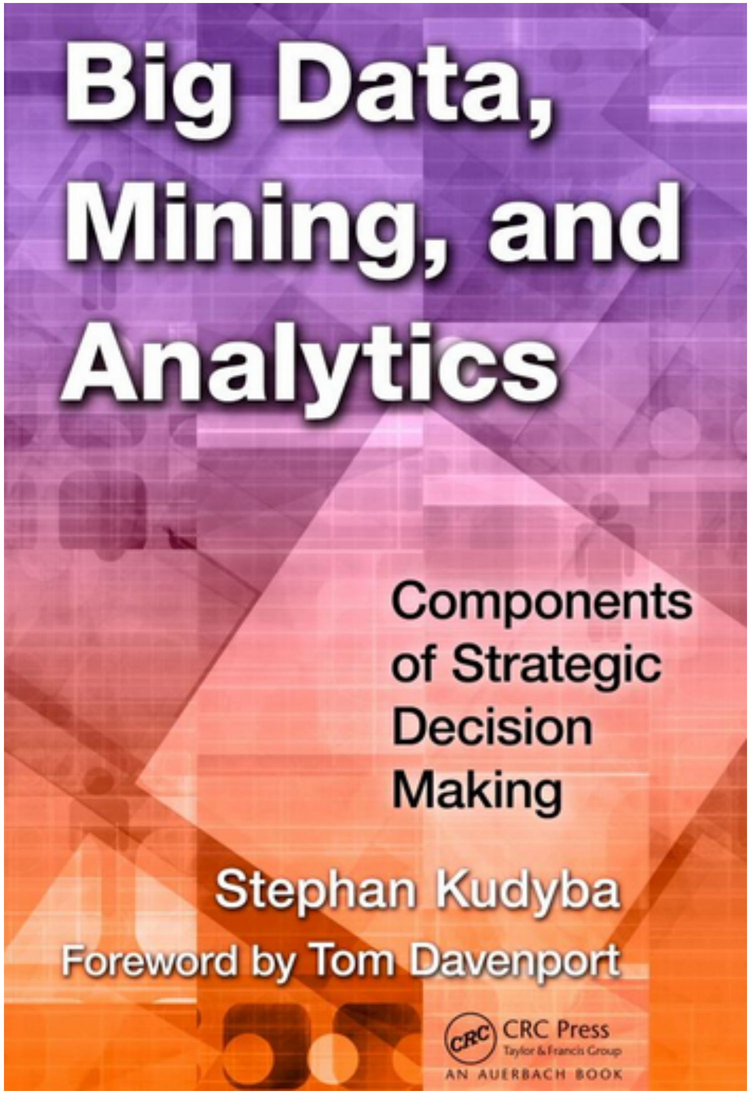


Statistics
Management Science & Information Systems
Artificial Intelligence
DATA MINING (Knowledge Discovery)
Database Management & Data Warehousing
Machine Learning & Pattern Recognition
Information Visualization

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

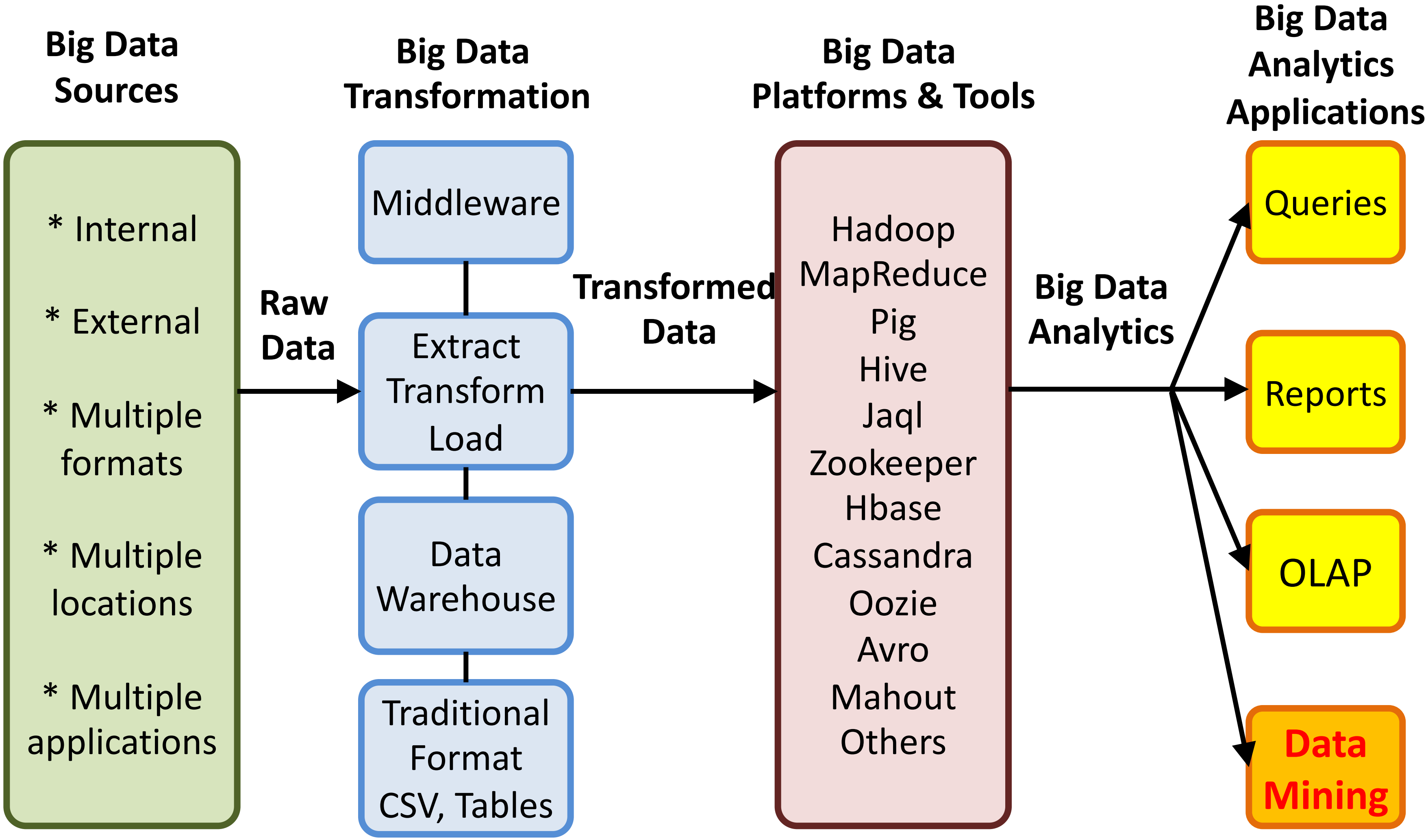
Stephan Kudyba (2014),

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


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

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

Architecture of Big Data Analytics



Big Data Sources

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

Raw Data

Big Data Transformation

- Middleware
- Extract Transform Load
- Data Warehouse
- Traditional Format CSV, Tables

Transformed Data

Big Data Platforms & Tools

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

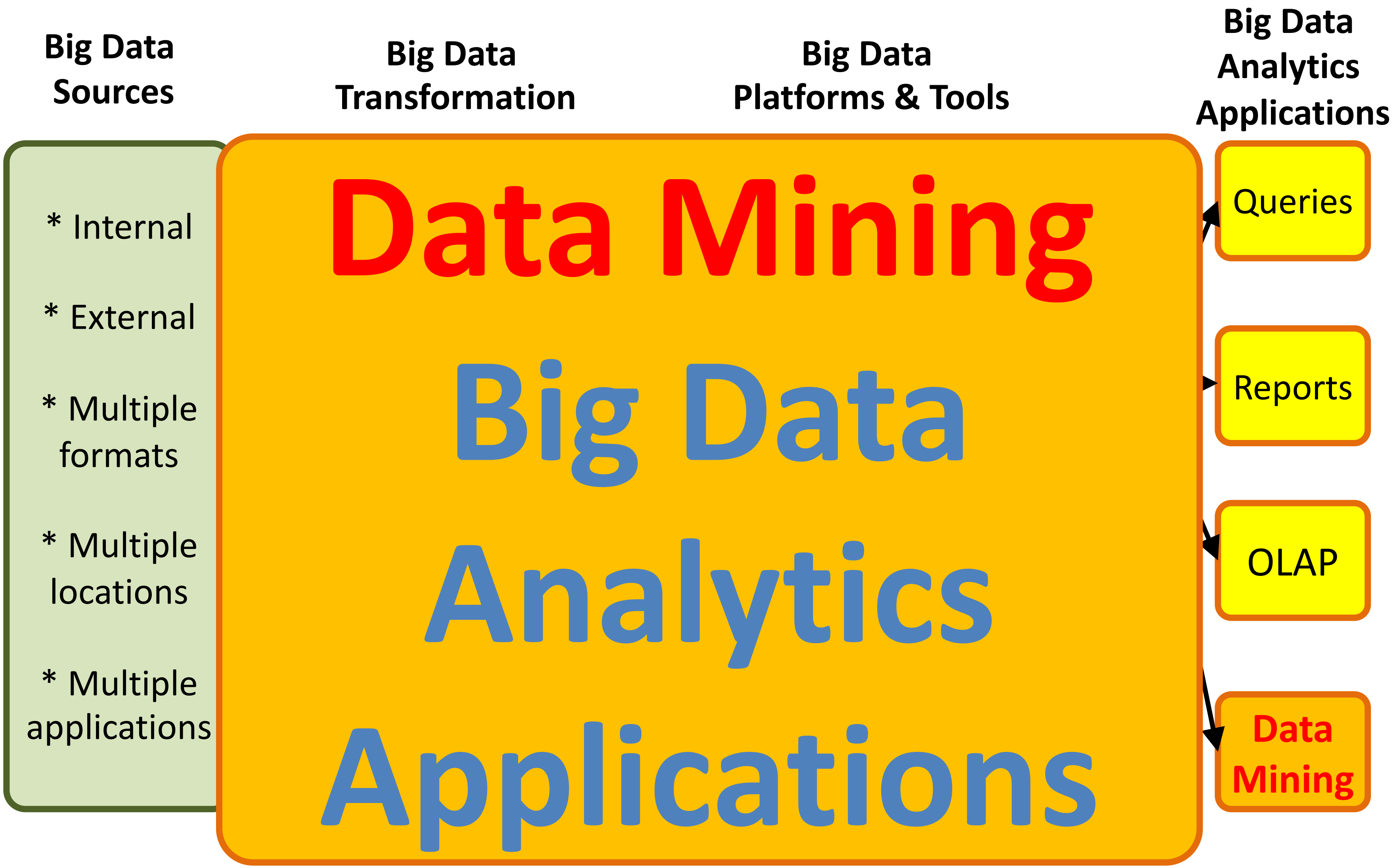
Big Data Analytics

Big Data Analytics Applications

- Queries
- Reports
- OLAP
- Data Mining

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

Architecture of Big Data Analytics



Big Data Sources

Big Data Transformation

Big Data Platforms & Tools

Big Data Analytics Applications

* Internal

* External

* Multiple formats

* Multiple locations

* Multiple applications

Data Mining

Big Data Analytics Applications

Queries

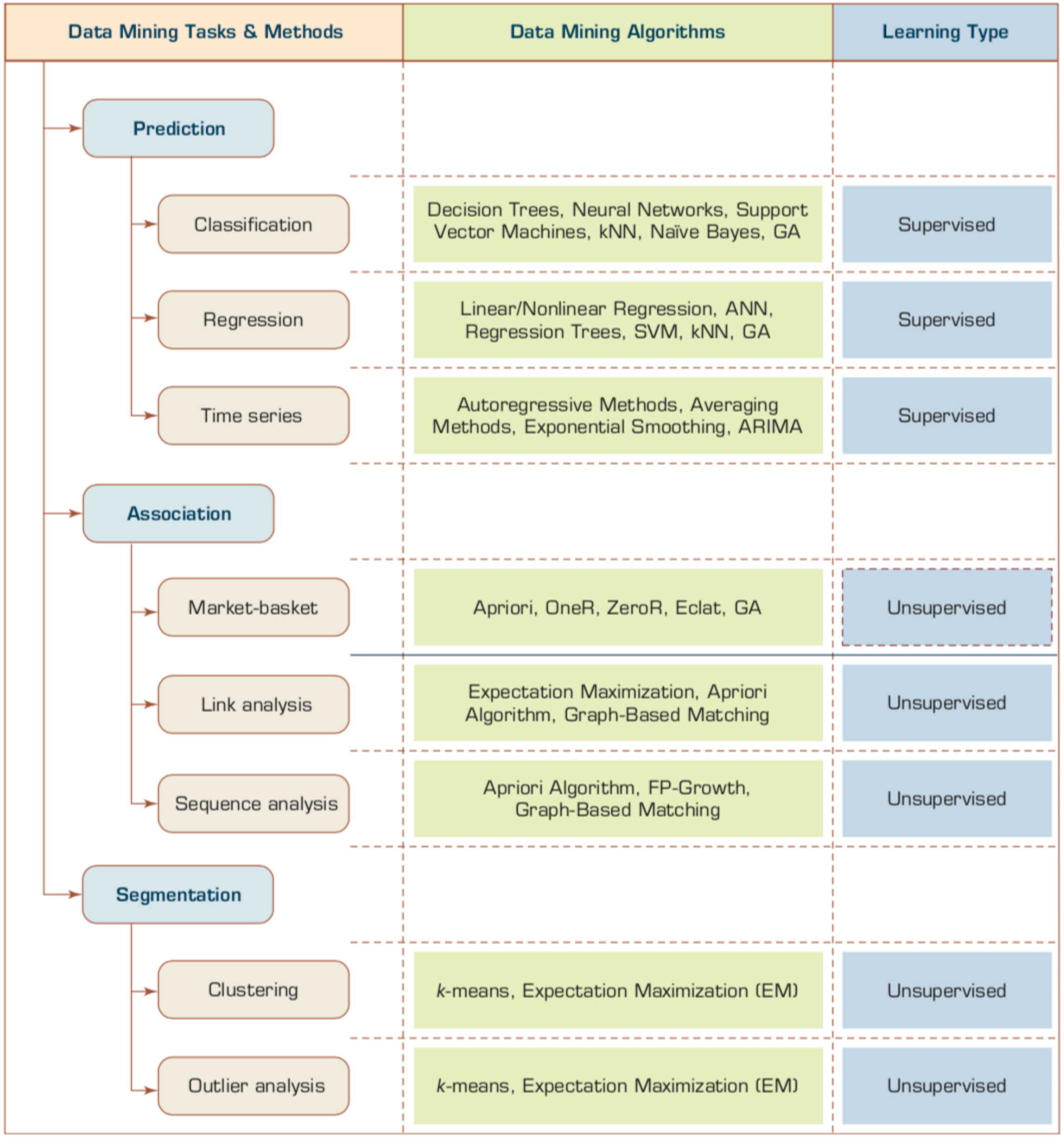
Reports

OLAP

Data Mining

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

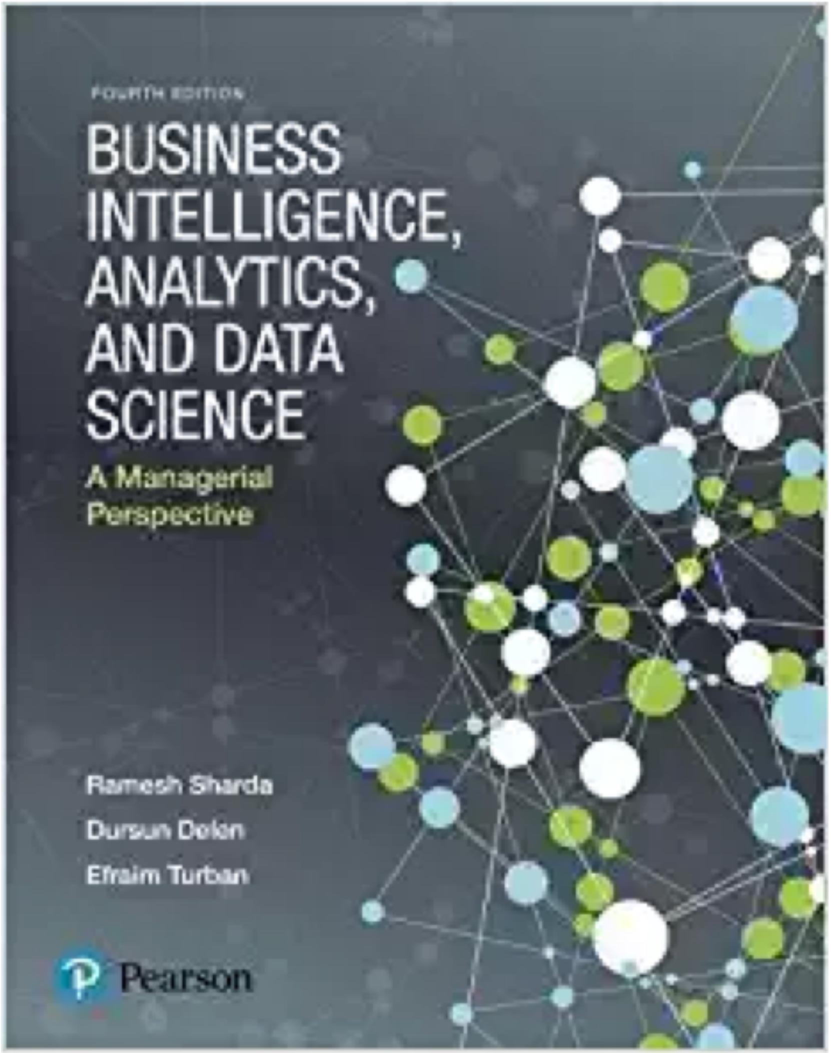
Data Mining Tasks & Methods



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

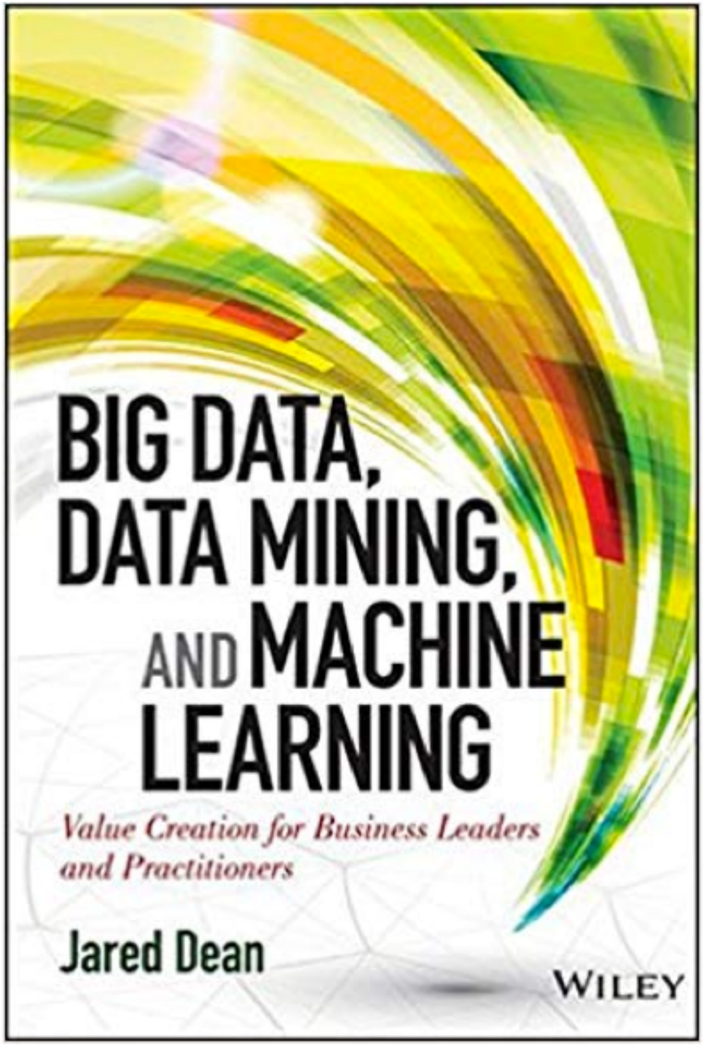
Source: Ramesh Sharda, Dursun Delen, and Efraim Turban (2017), Business Intelligence, Analytics, and Data Science: A Managerial Perspective, 4th Edition, Pearson

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



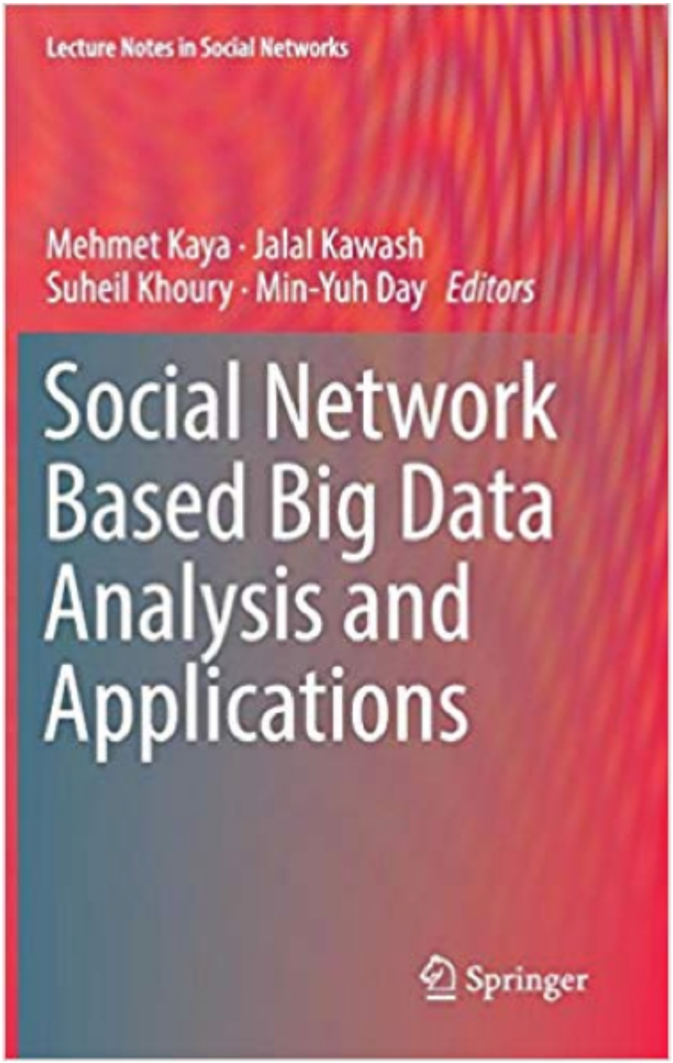
Source: https://www.amazon.com/Business-Intelligence-Analytics-Data-Science/dp/0134633288

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

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



Lecture Notes in Social Networks

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

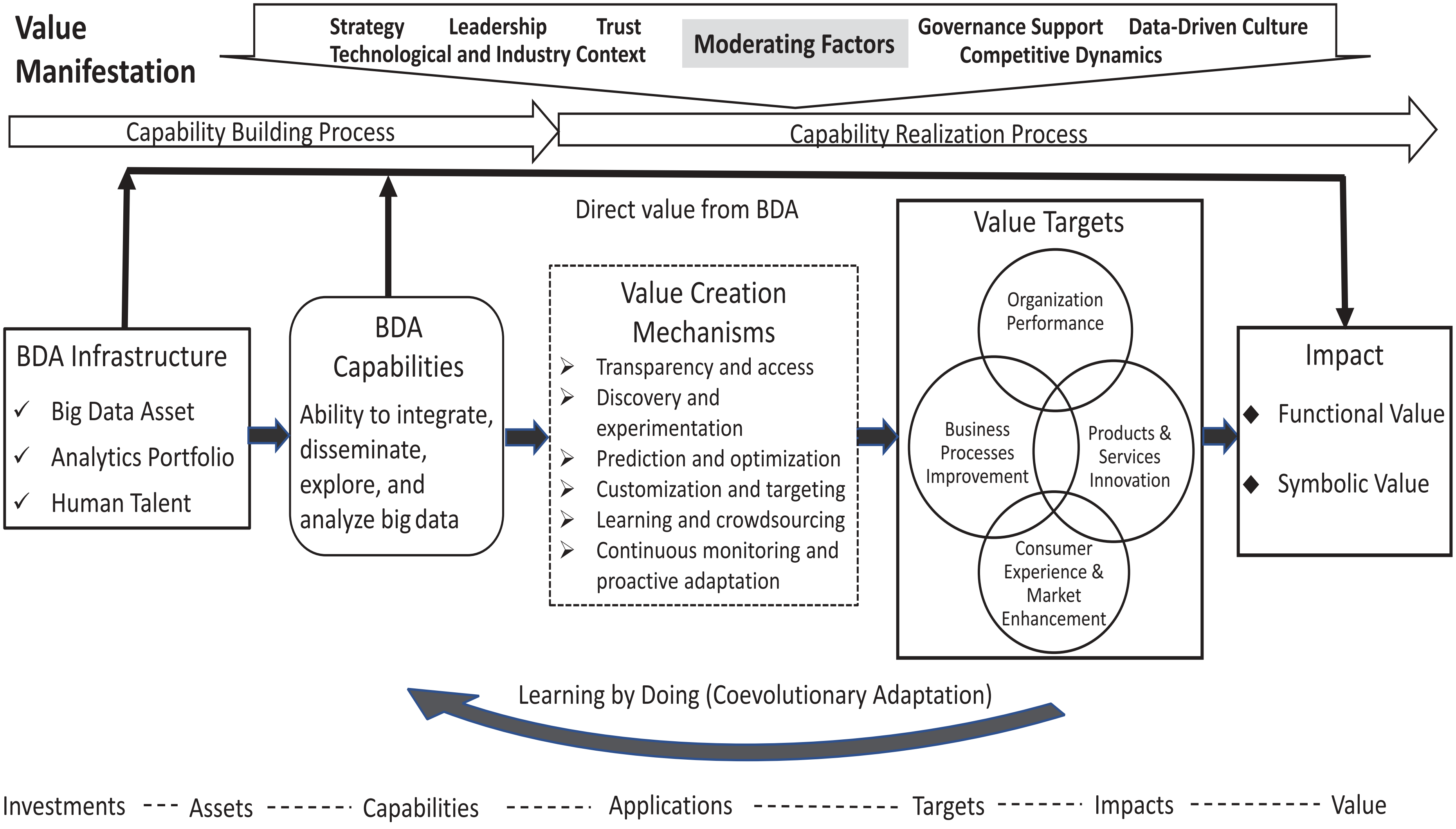
Social Network Based Big Data Analysis and Applications

Springer

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

Value Creation by Big Data Analytics

(Grover et al., 2018)



Value Manifestation
Strategy
Leadership
Trust
Technological and Industry Context
Moderating Factors
Governance Support
Data-Driven Culture
Competitive Dynamics
Capability Building Process
Capability Realization Process
Direct value from BDA
BDA Infrastructure
✓ Big Data Asset
✓ Analytics Portfolio
✓ Human Talent
BDA Capabilities
Ability to integrate, disseminate, explore, and analyze big data
Value Creation Mechanisms
➢ Transparency and access
➢ Discovery and experimentation
➢ Prediction and optimization
➢ Customization and targeting
➢ Learning and crowdsourcing
➢ Continuous monitoring and proactive adaptation
Value Targets
Organization Performance
Business Processes Improvement
Products & Services Innovation
Consumer Experience & Market Enhancement
Impact
◆ Functional Value
◆ Symbolic Value
Learning by Doing (Coevolutionary Adaptation)
Investments --- Assets ------ Capabilities ------ Applications ---------- Targets ------ Impacts --------- Value

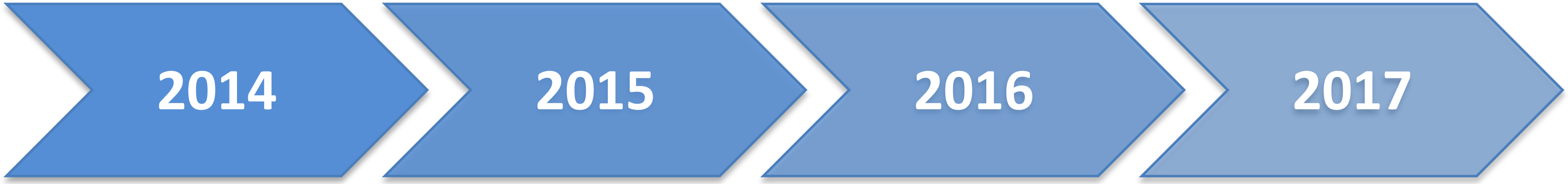
Source: Varun Grover, Roger HL Chiang, Ting-Peng Liang, and Dongsong Zhang (2018), "Creating Strategic Business Value from Big Data Analytics: A Research Framework", Journal of Management Information Systems, 35, no. 2, pp. 388-423.

Research Landscape of Business Intelligence and Big Data Analytics: A bibliometrics study

- A bibliometric analysis on Big Data and Business Intelligence from 1990 to 2016.
- Big Data papers grow much faster than Business Intelligence papers
- Computer Science and information systems are two core disciplines.
- Most influential papers are identified and a research framework is proposed.

Source: Ting-Peng Liang and Yu-Hsi Liu (2018), "Research Landscape of Business Intelligence and Big Data analytics: A bibliometrics study”,

Evolution of top keywords in “BD & BI” publications



- Management
- Text Mining
- Data Mining
- Data Science

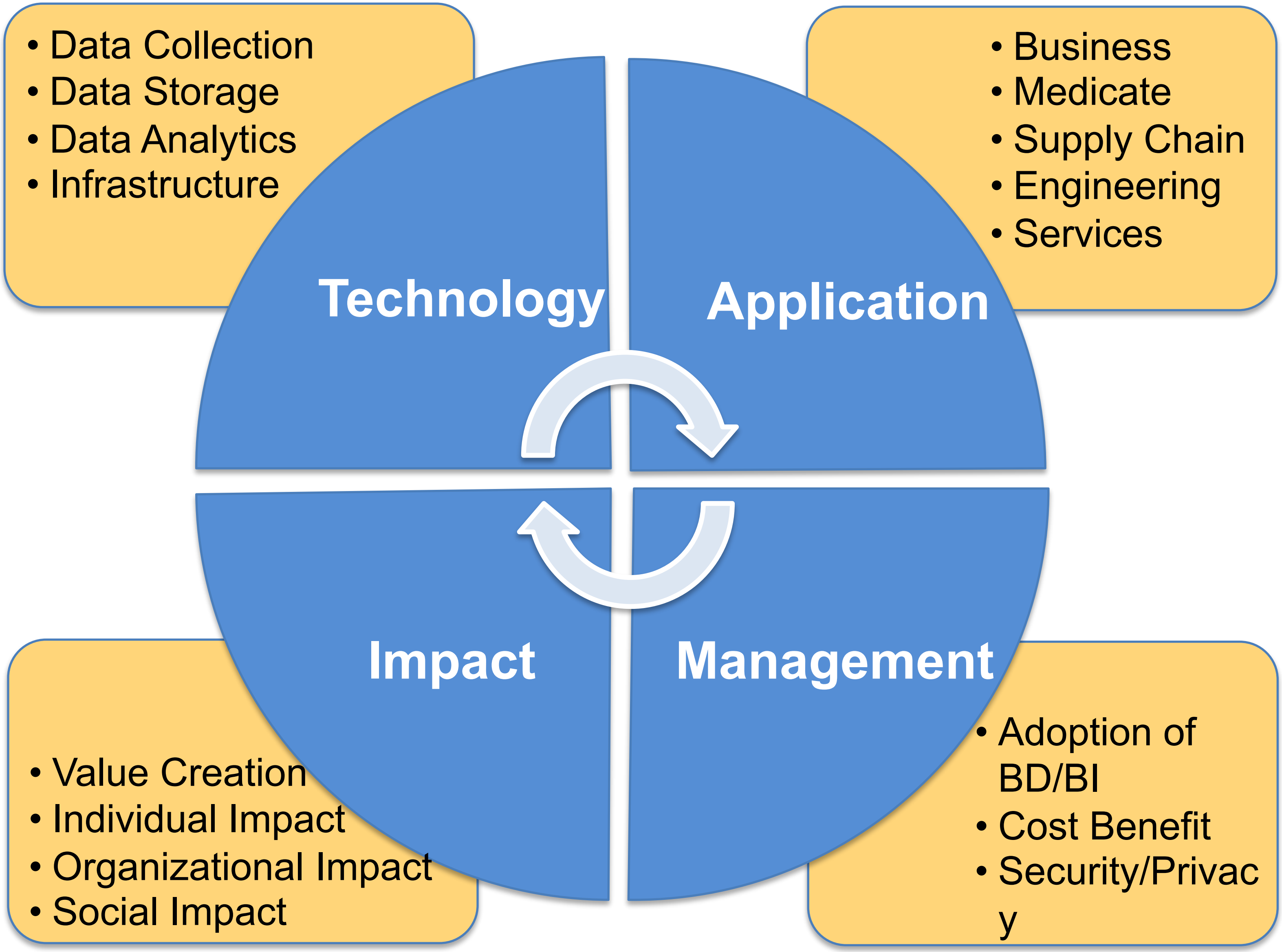
- Big Data Analytics
- Social Media
- Business Analytics
- Information System

- Cloud Computing
- Data Warehouse

- Knowledge Management

Source: Ting-Peng Liang and Yu-Hsi Liu (2018), "Research Landscape of Business Intelligence and Big Data analytics: A bibliometrics study”,
Expert Systems with Applications, Volume 111, 30, 2018, pp. 2-10

Framework for BD and BI Research



- Data Collection
- Data Storage
- Data Analytics
- Infrastructure

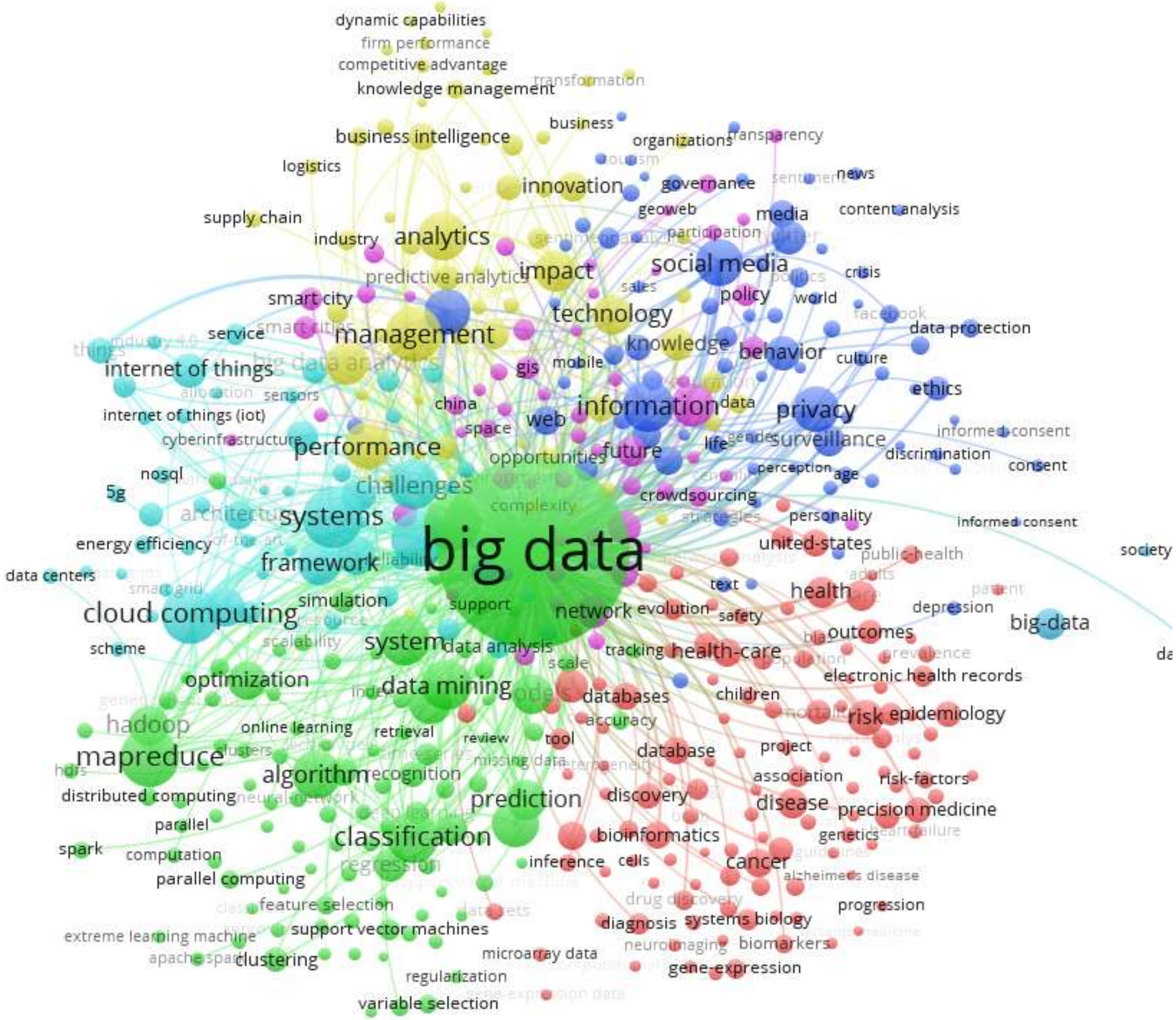
- Business
- Medicate
- Supply Chain
- Engineering
- Services

- Value Creation
- Individual Impact
- Organizational Impact
- Social Impact

- Adoption of BD/BI
- Cost Benefit
- Security/Privacy
- Human Resource

Source: Ting-Peng Liang and Yu-Hsi Liu (2018), "Research Landscape of Business Intelligence and Big Data analytics: A bibliometrics study",
Expert Systems with Applications, Volume 111, 30, 2018, pp. 2-10

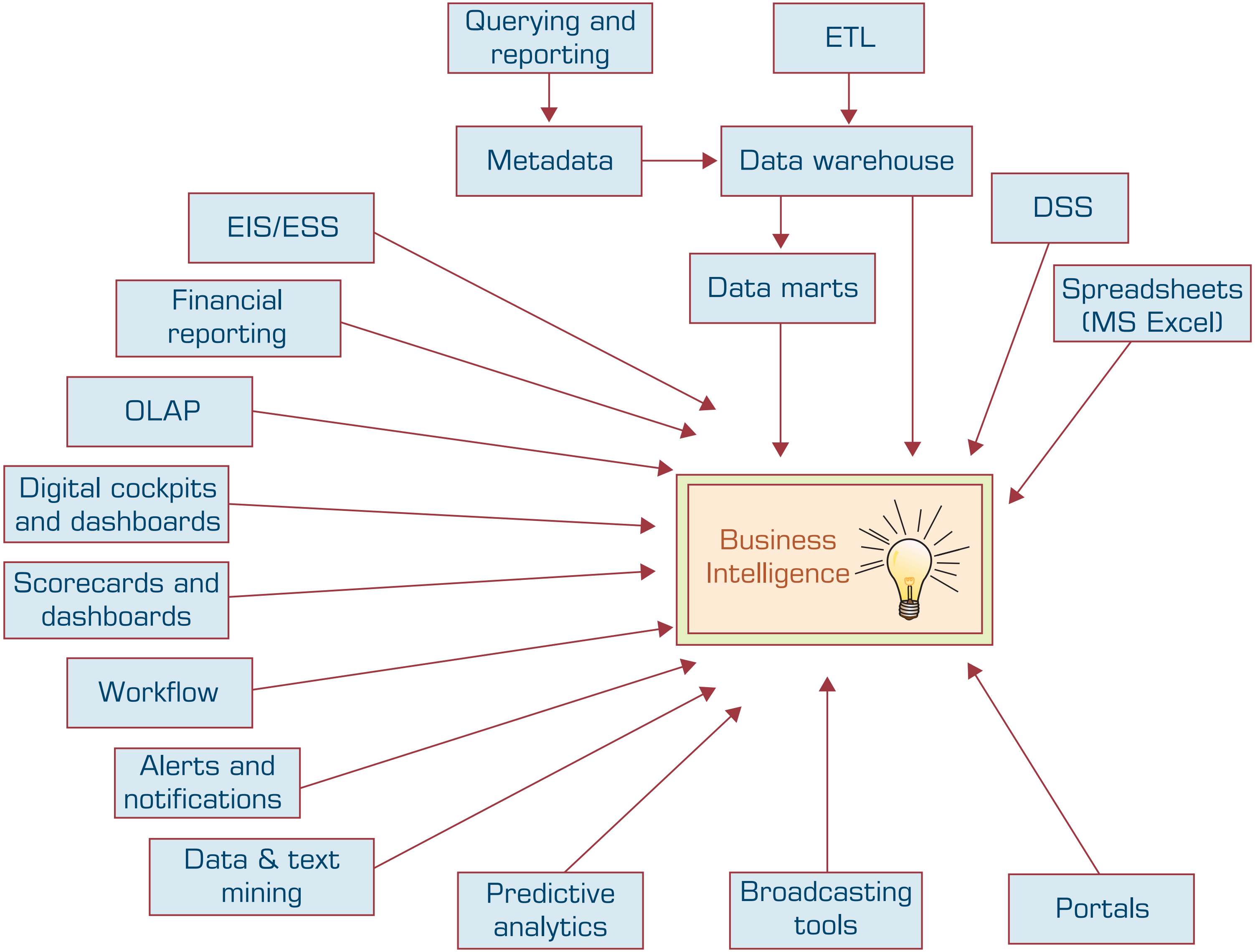
Business Intelligence and Big Data analytics



Source: Ting-Peng Liang and Yu-Hsi Liu (2018), "Research Landscape of Business Intelligence and Big Data analytics: A bibliometrics study",

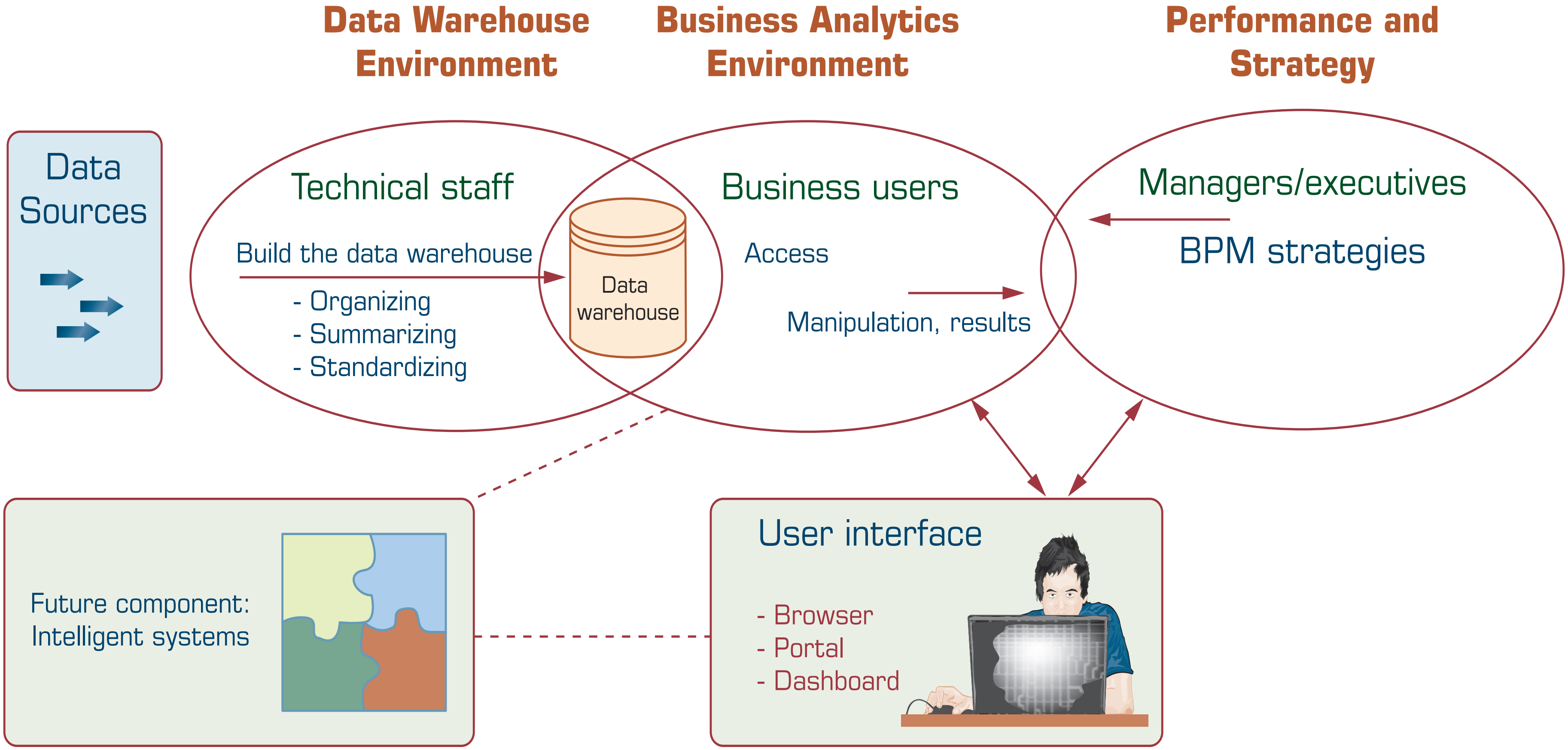
Expert Systems with Applications, Volume 111, 30, 2018, pp. 2-10

Evolution of Business Intelligence (BI)



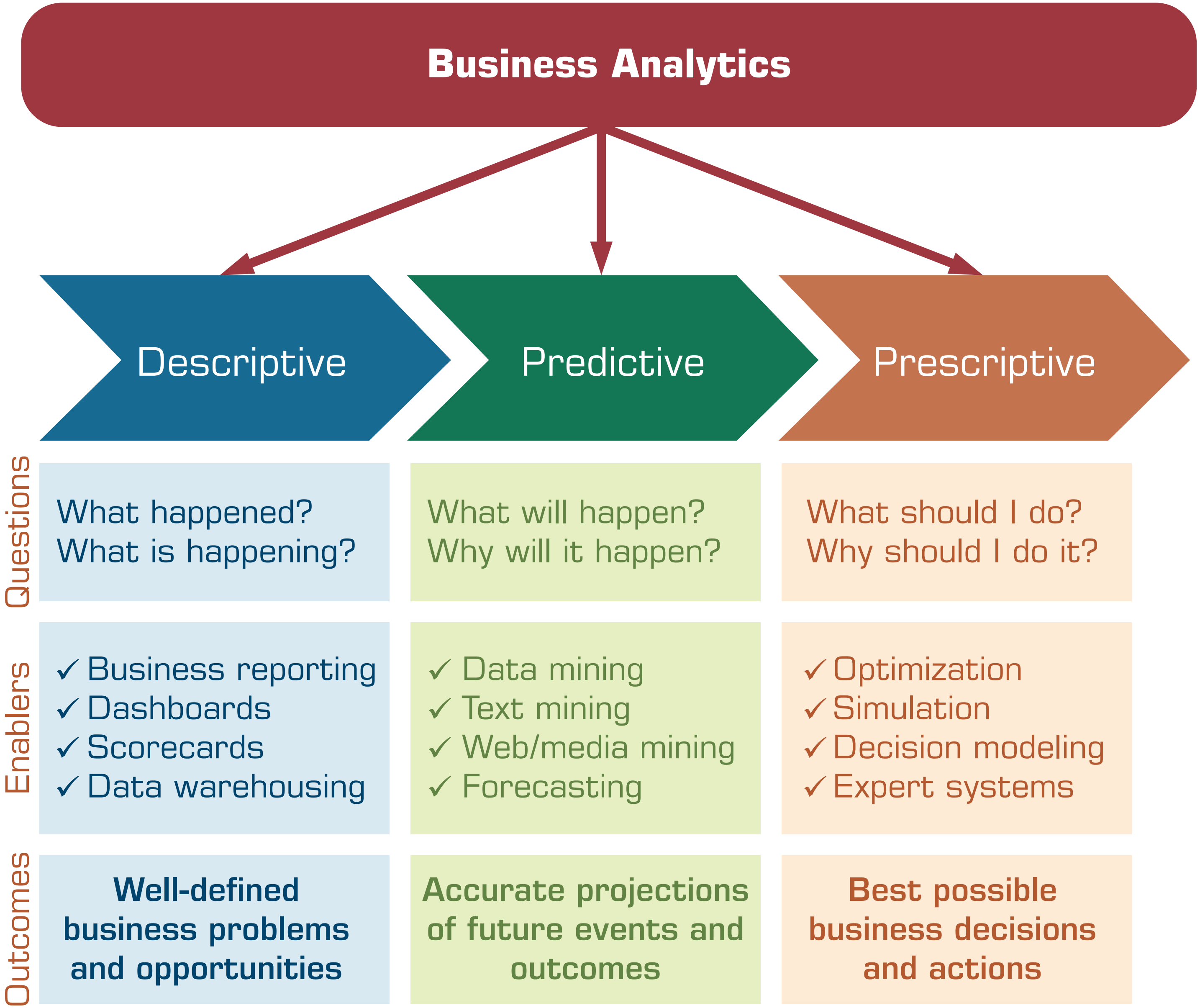
Source: Ramesh Sharda, Dursun Delen, and Efraim Turban (2017), Business Intelligence, Analytics, and Data Science: A Managerial Perspective, 4th Edition, Pearson

A High-Level Architecture of BI



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

Three Types of Analytics

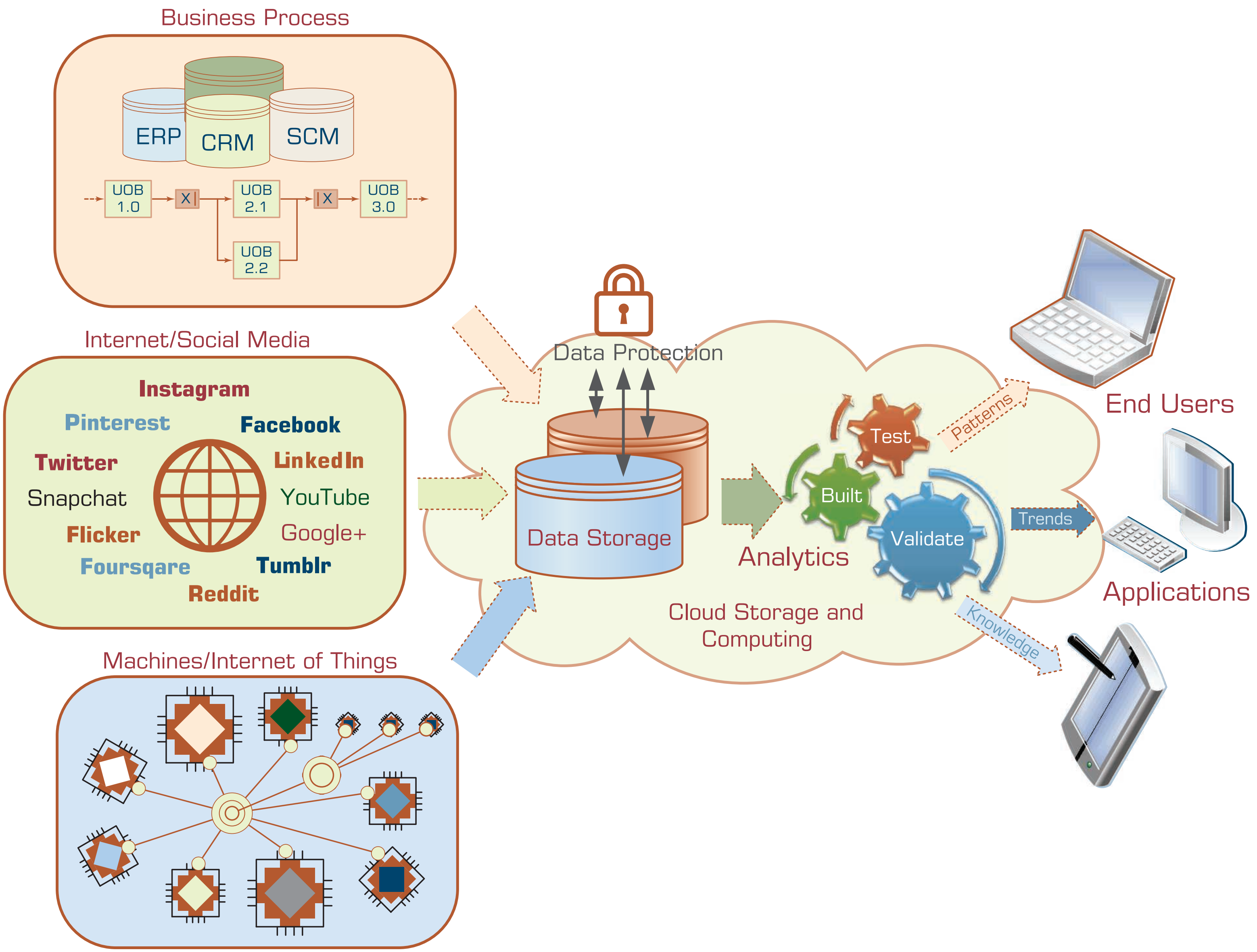


Business Analytics

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

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

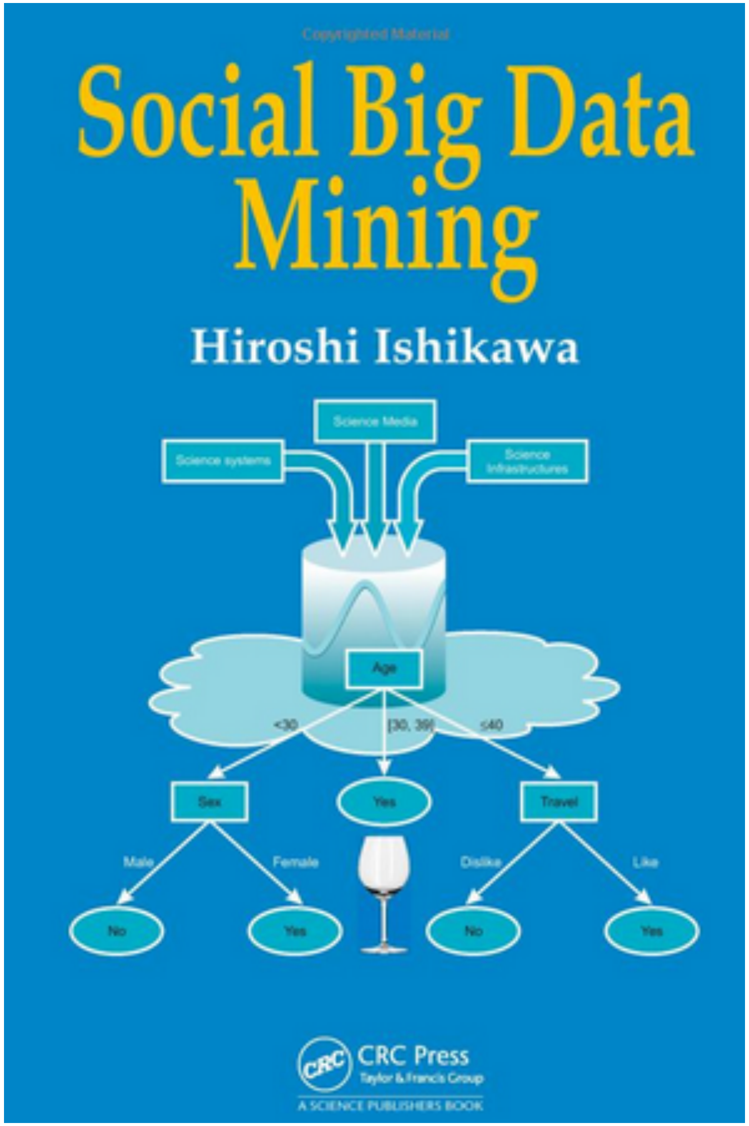
A Data to Knowledge Continuum



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

Social Big Data Mining

(Hiroshi Ishikawa, 2015)



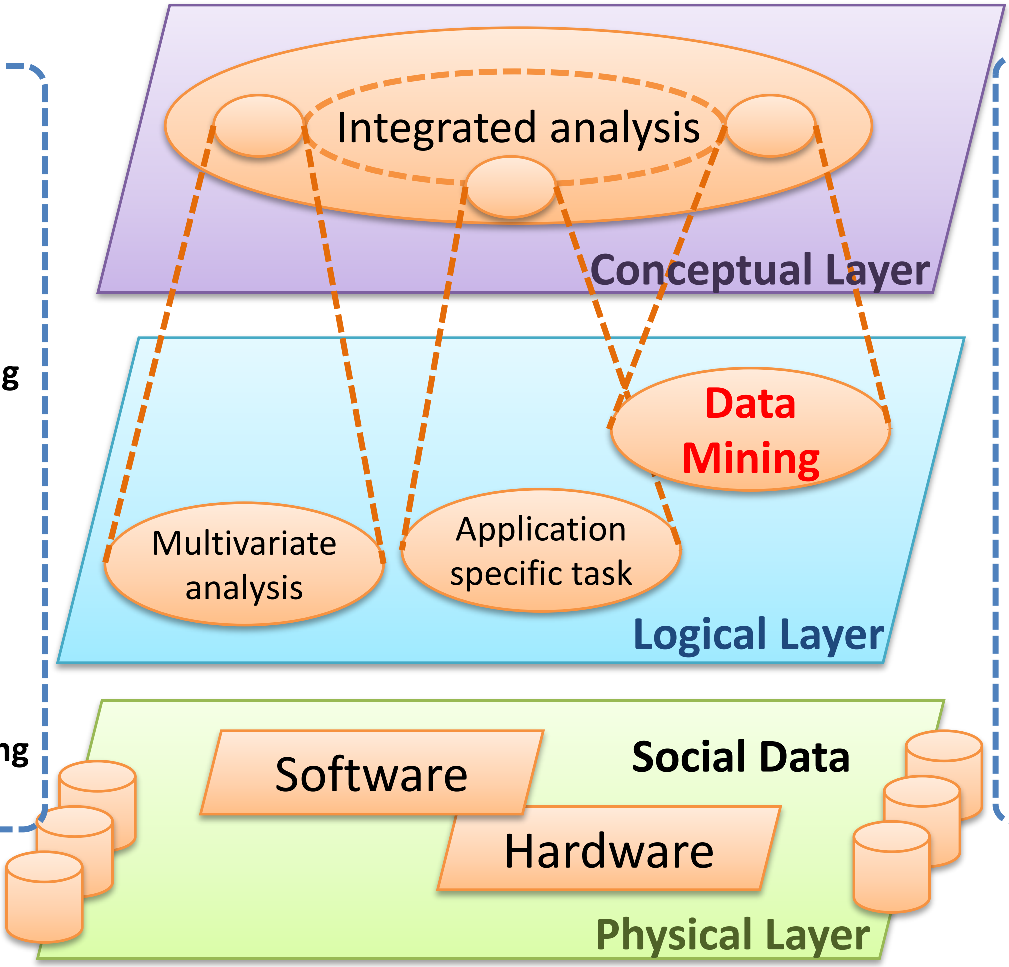
Source: http://www.amazon.com/Social-Data-Mining-Hiroshi-Ishikawa/dp/149871093X

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**

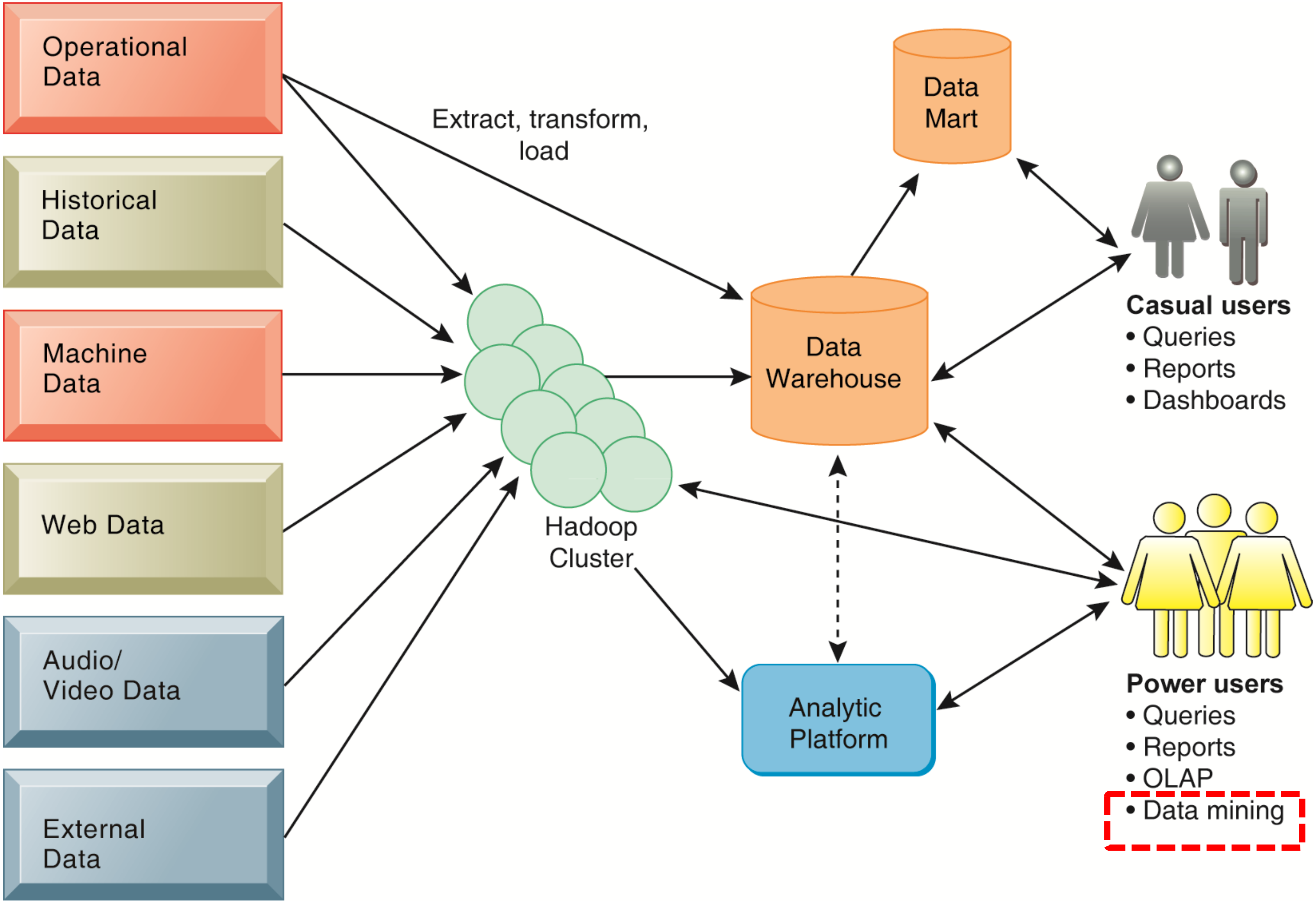


Analysts

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

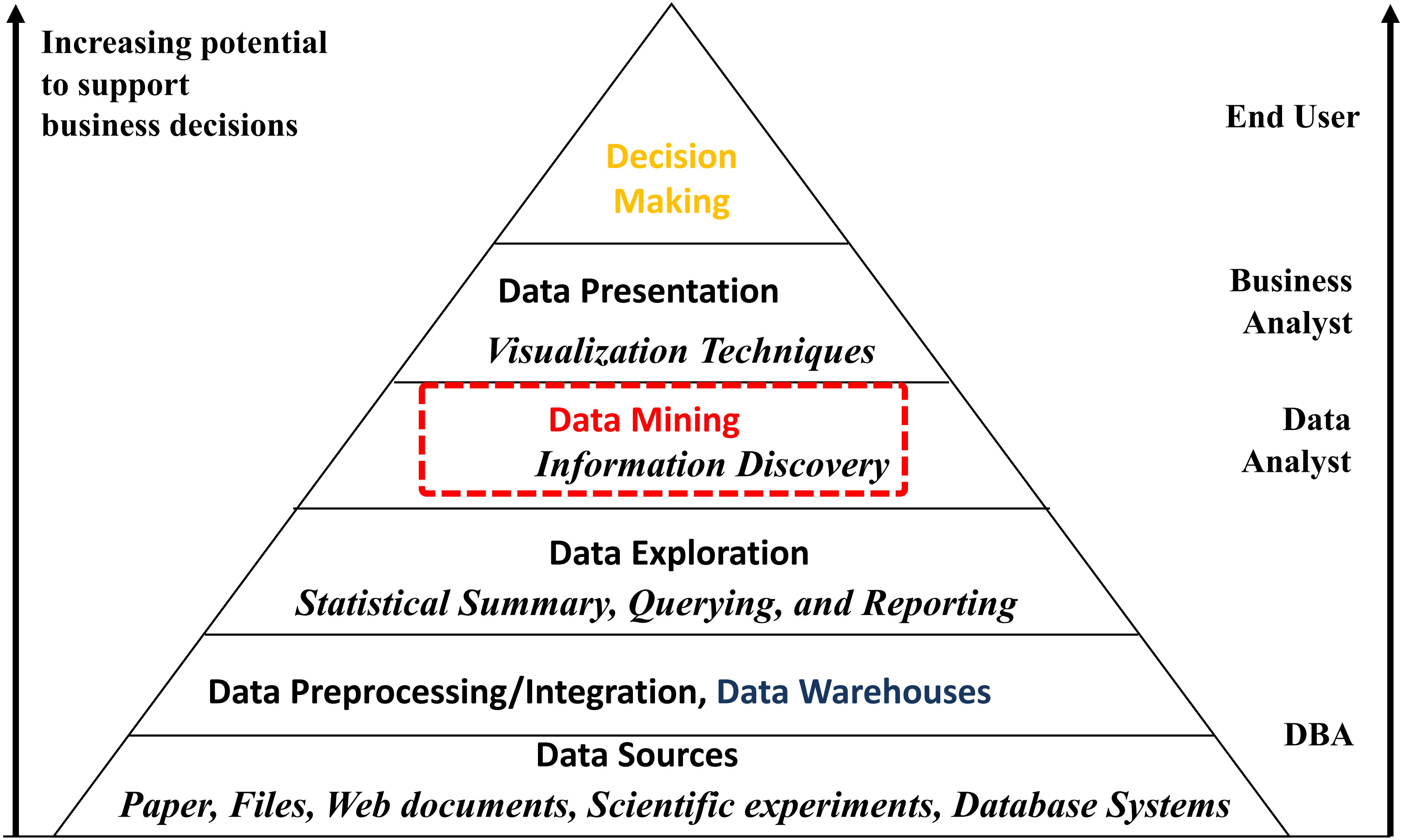
Source: Hiroshi Ishikawa (2015), Social Big Data Mining, CRC Press

Business Intelligence (BI) Infrastructure



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

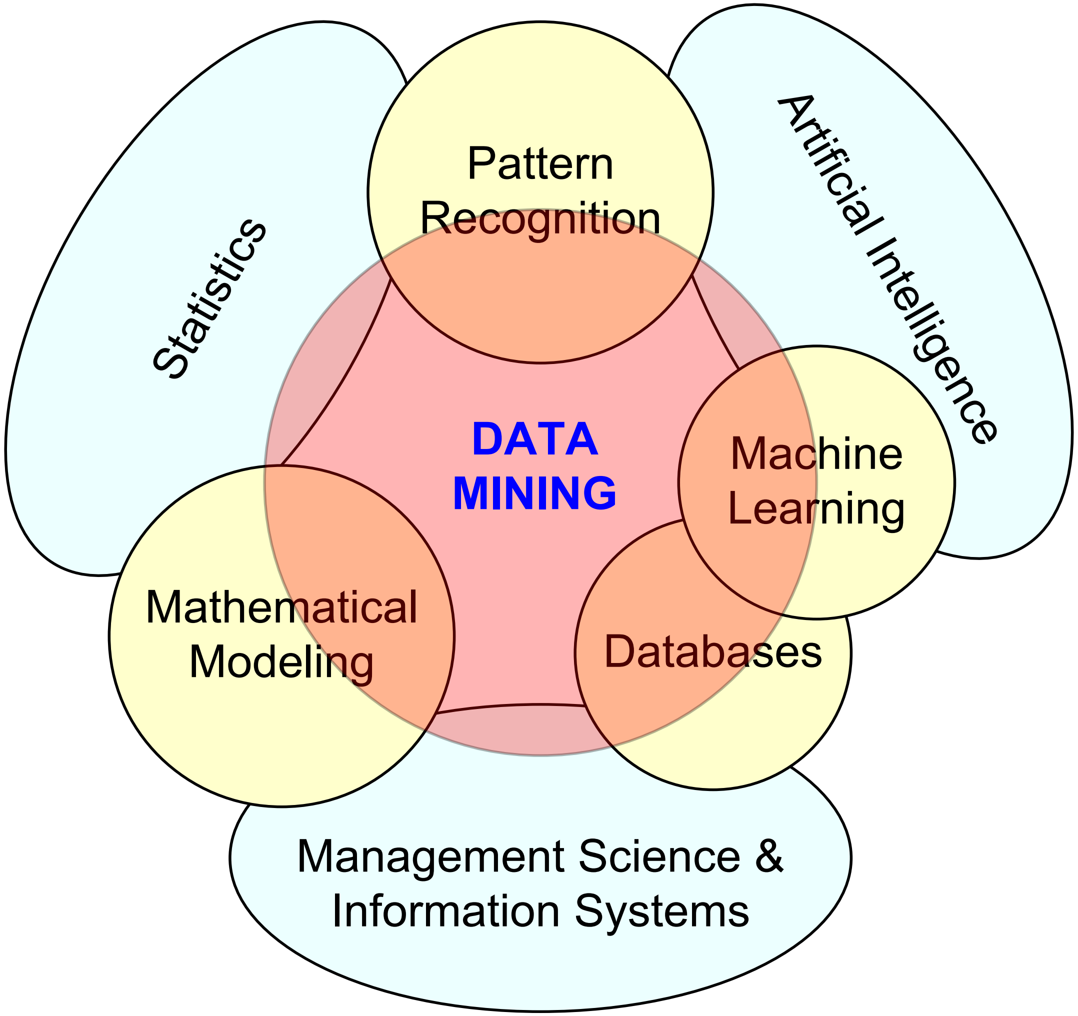
Business Intelligence and Data Mining



Increasing potential to support business decisions
Decision Making
Data Presentation
Visualization Techniques
Data Mining
Information Discovery
Data Exploration
Statistical Summary, Querying, and Reporting
Data Preprocessing/Integration, Data Warehouses
Data Sources
Paper, Files, Web documents, Scientific experiments, Database Systems
End User
Business Analyst
Data Analyst
DBA

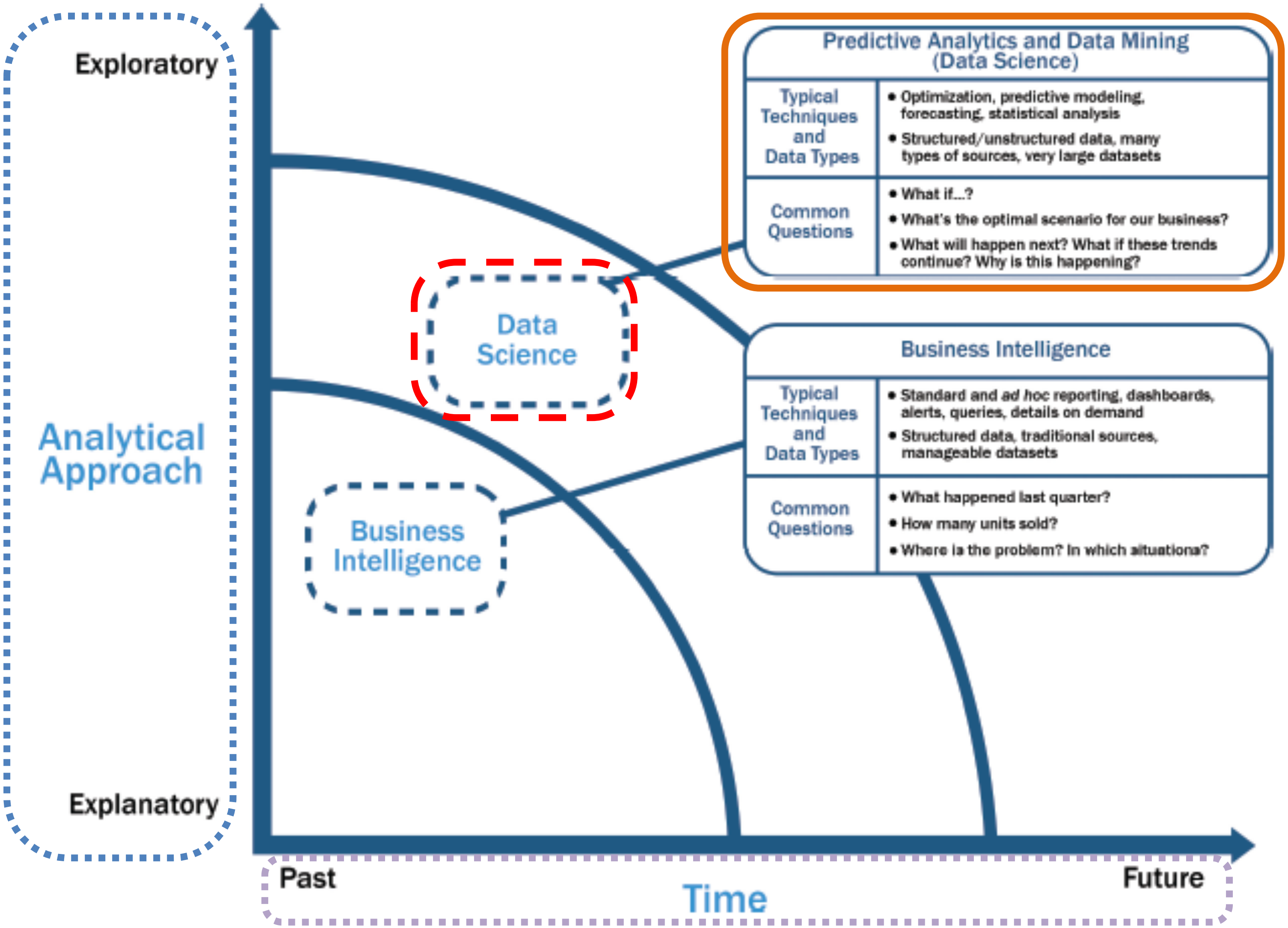
Source: Jiawei Han and Micheline Kamber (2006), Data Mining: Concepts and Techniques, Second Edition, Elsevier

Data Mining at the Intersection of Many Disciplines



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

Data Science and Business Intelligence

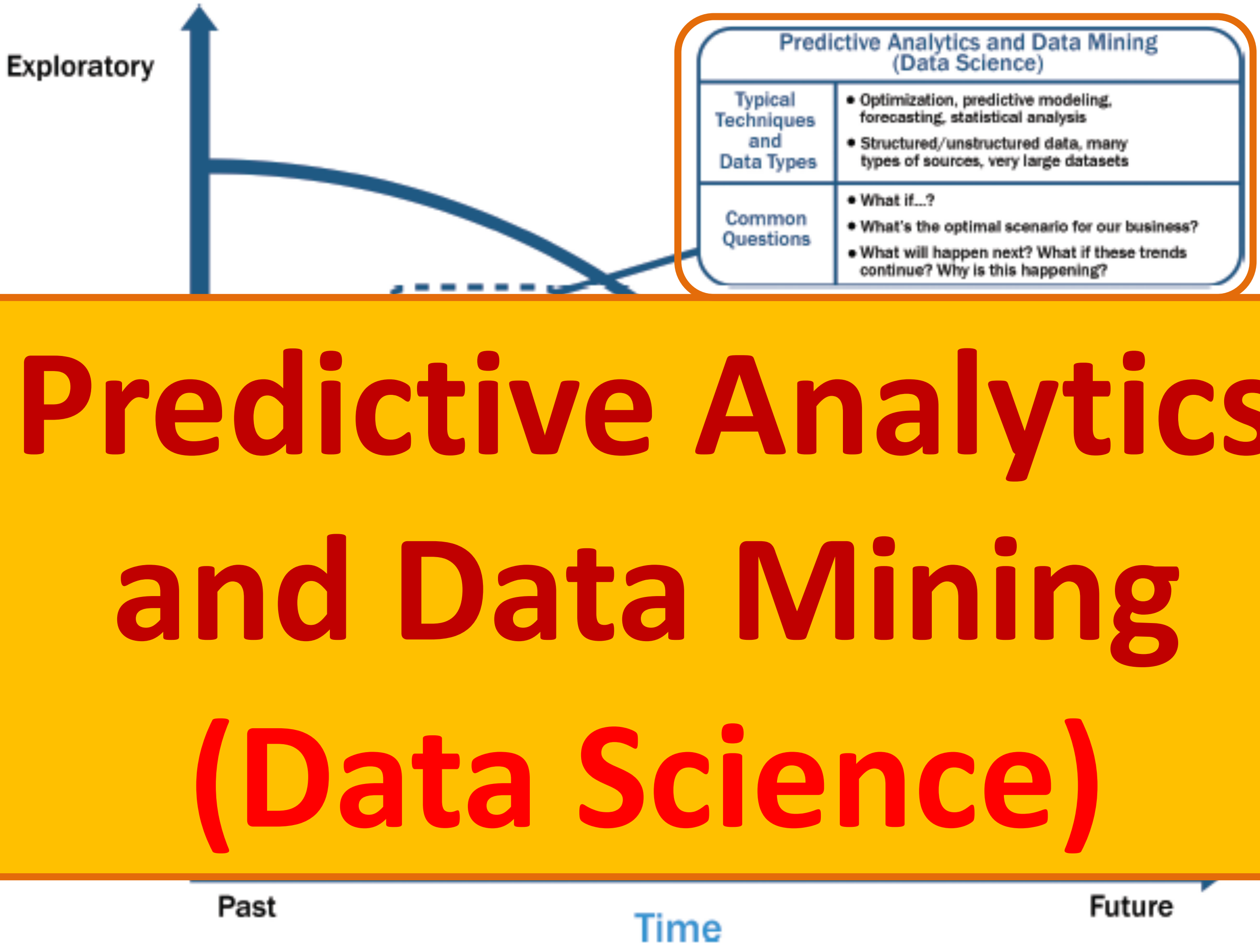


Predictive Analytics and Data Mining (Data Science)	
Typical Techniques and Data Types	• Optimization, predictive modeling, forecasting, statistical analysis • Structured/unstructured data, many types of sources, very large datasets
Common Questions	• What if...? • What's the optimal scenario for our business? • What will happen next? What if these trends continue? Why is this happening?

Business Intelligence	
Typical Techniques and Data Types	• Standard and *ad hoc* reporting, dashboards, alerts, queries, details on demand • Structured data, traditional sources, manageable datasets
Common Questions	• What happened last quarter? • How many units sold? • Where is the problem? In which situations?

Source: EMC Education Services, Data Science and Big Data Analytics: Discovering, Analyzing, Visualizing and Presenting Data, Wiley, 2015

Data Science and Business Intelligence



Predictive Analytics and Data Mining (Data Science)	
Typical Techniques and Data Types	• Optimization, predictive modeling, forecasting, statistical analysis • Structured/unstructured data, many types of sources, very large datasets
Common Questions	• What if...? • What's the optimal scenario for our business? • What will happen next? What if these trends continue? Why is this happening?

Predictive Analytics and Data Mining (Data Science)

Source: EMC Education Services, Data Science and Big Data Analytics: Discovering, Analyzing, Visualizing and Presenting Data, Wiley, 2015

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 countinue?
Why is this happening?

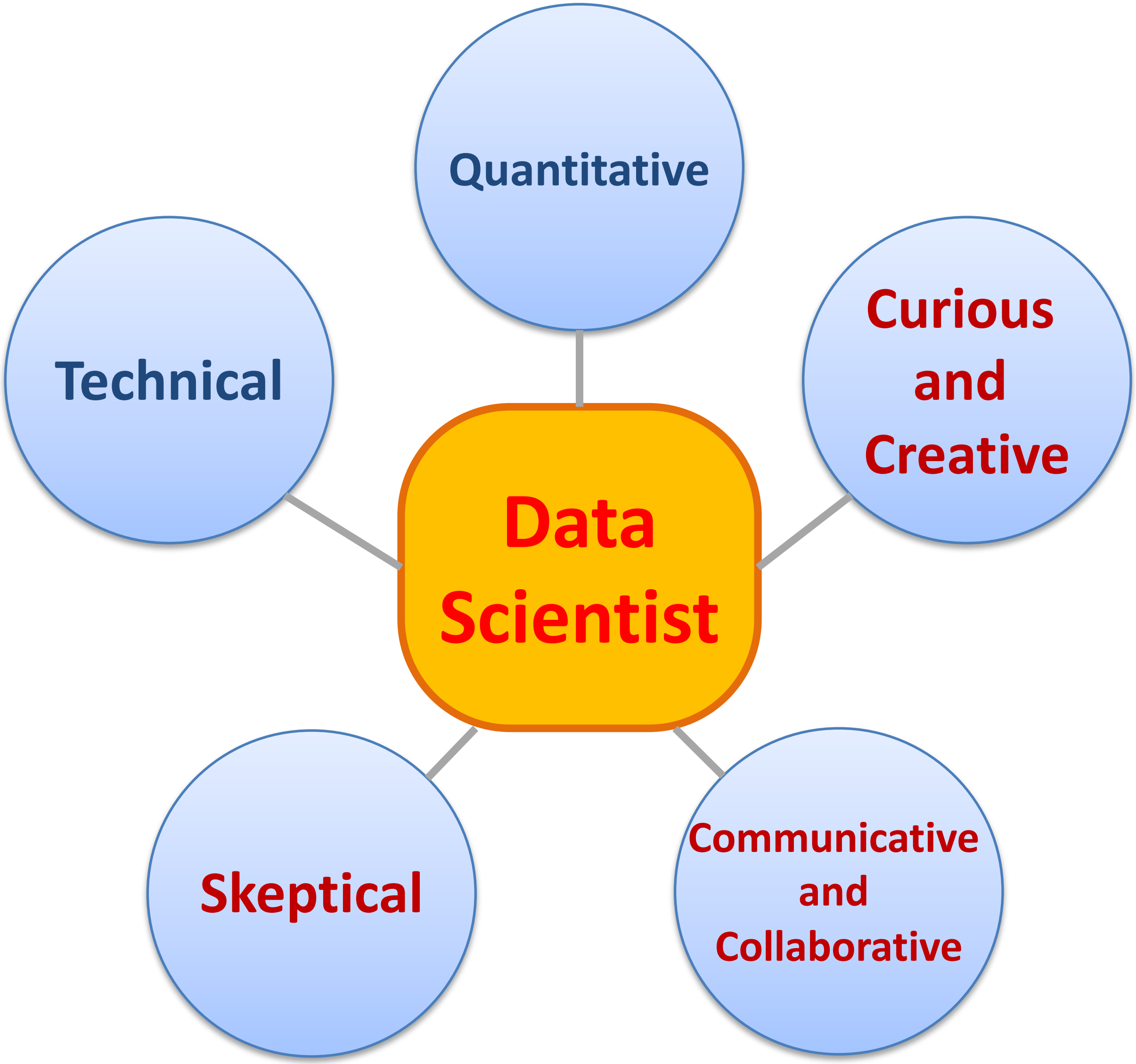
Source: EMC Education Services, Data Science and Big Data Analytics: Discovering, Analyzing, Visualizing and Presenting Data, Wiley, 2015

Profile of a Data Scientist

- **Quantitative**
  - mathematics or statistics
- **Technical**
  - software engineering, machine learning, and programming skills
- **Skeptical mind-set** and **critical thinking**
- **Curious** and **creative**
- **Communicative** and **collaborative**

Source: EMC Education Services, Data Science and Big Data Analytics: Discovering, Analyzing, Visualizing and Presenting Data, Wiley, 2015

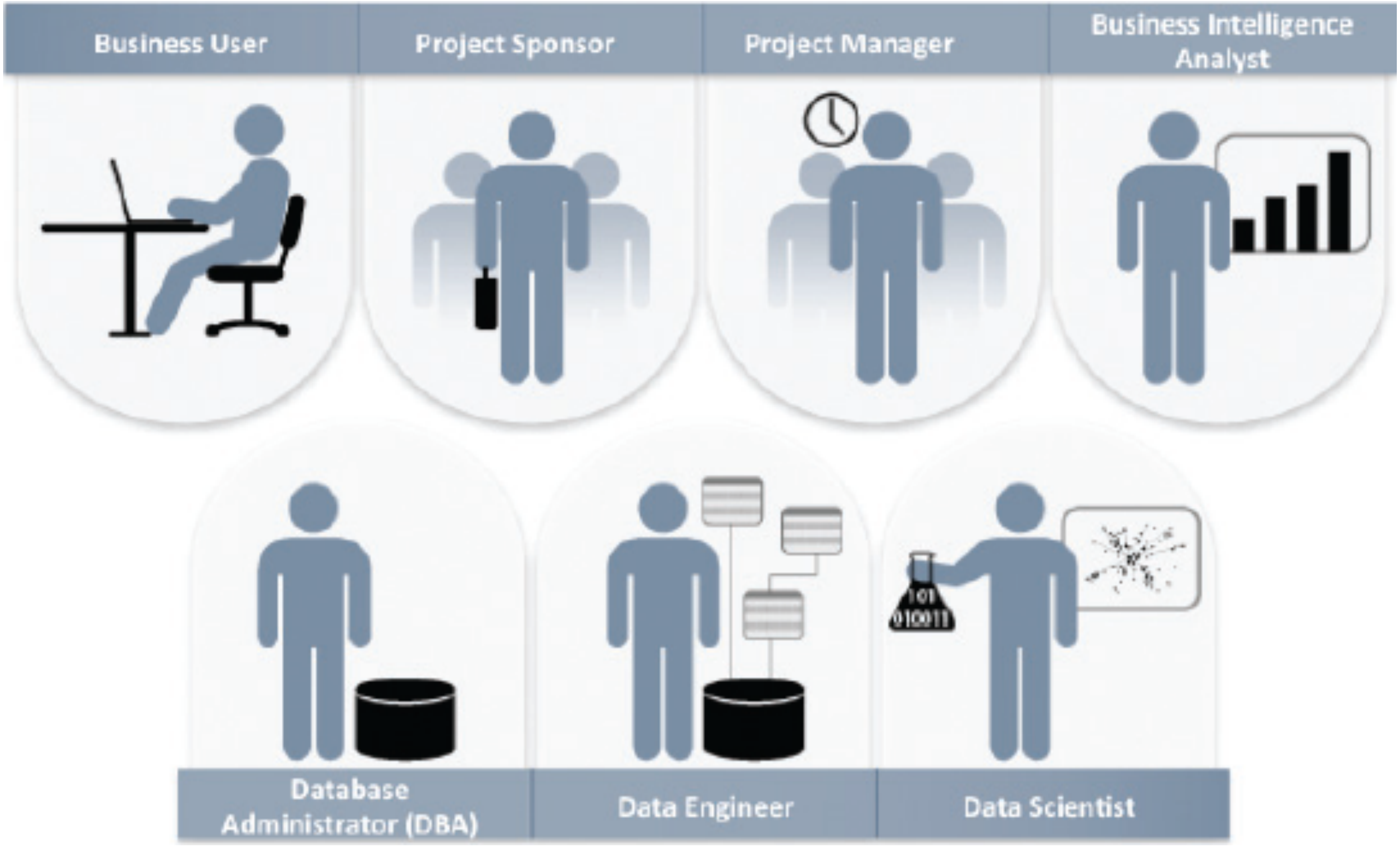
Data Scientist Profile



Source: EMC Education Services, Data Science and Big Data Analytics: Discovering, Analyzing, Visualizing and Presenting Data, Wiley, 2015

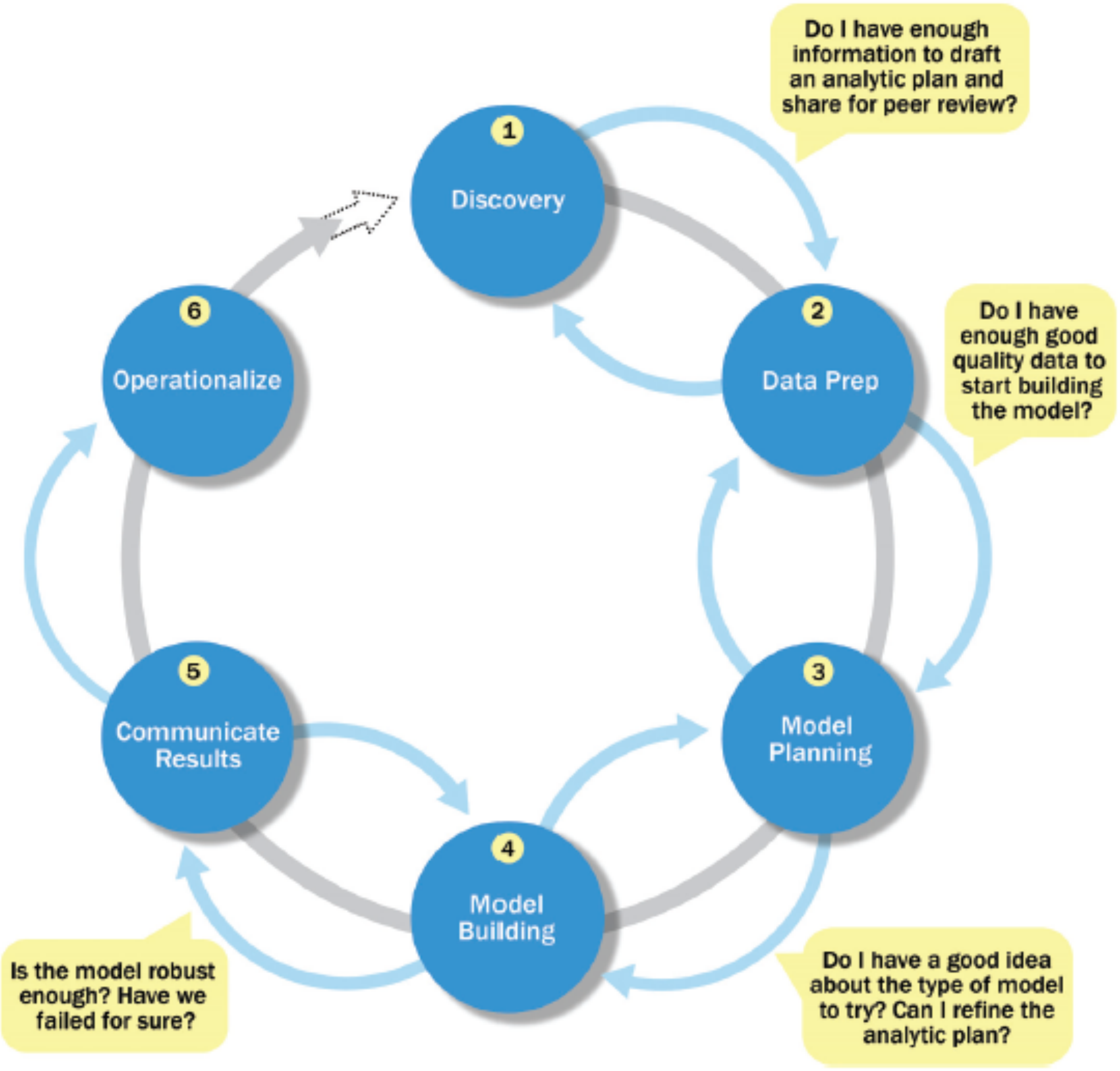
Big Data Analytics Lifecycle

Key Roles for a Successful Analytics Project



Source: EMC Education Services, Data Science and Big Data Analytics: Discovering, Analyzing, Visualizing and Presenting Data, Wiley, 2015

Overview of Data Analytics Lifecycle


1
Discovery
2
Data Prep
3
Model Planning
4
Model Building
5
Communicate Results
6
Operationalize
Do I have enough information to draft an analytic plan and share for peer review?
Do I have enough good quality data to start building the model?
Do I have a good idea about the type of model to try? Can I refine the analytic plan?
Is the model robust enough? Have we failed for sure?

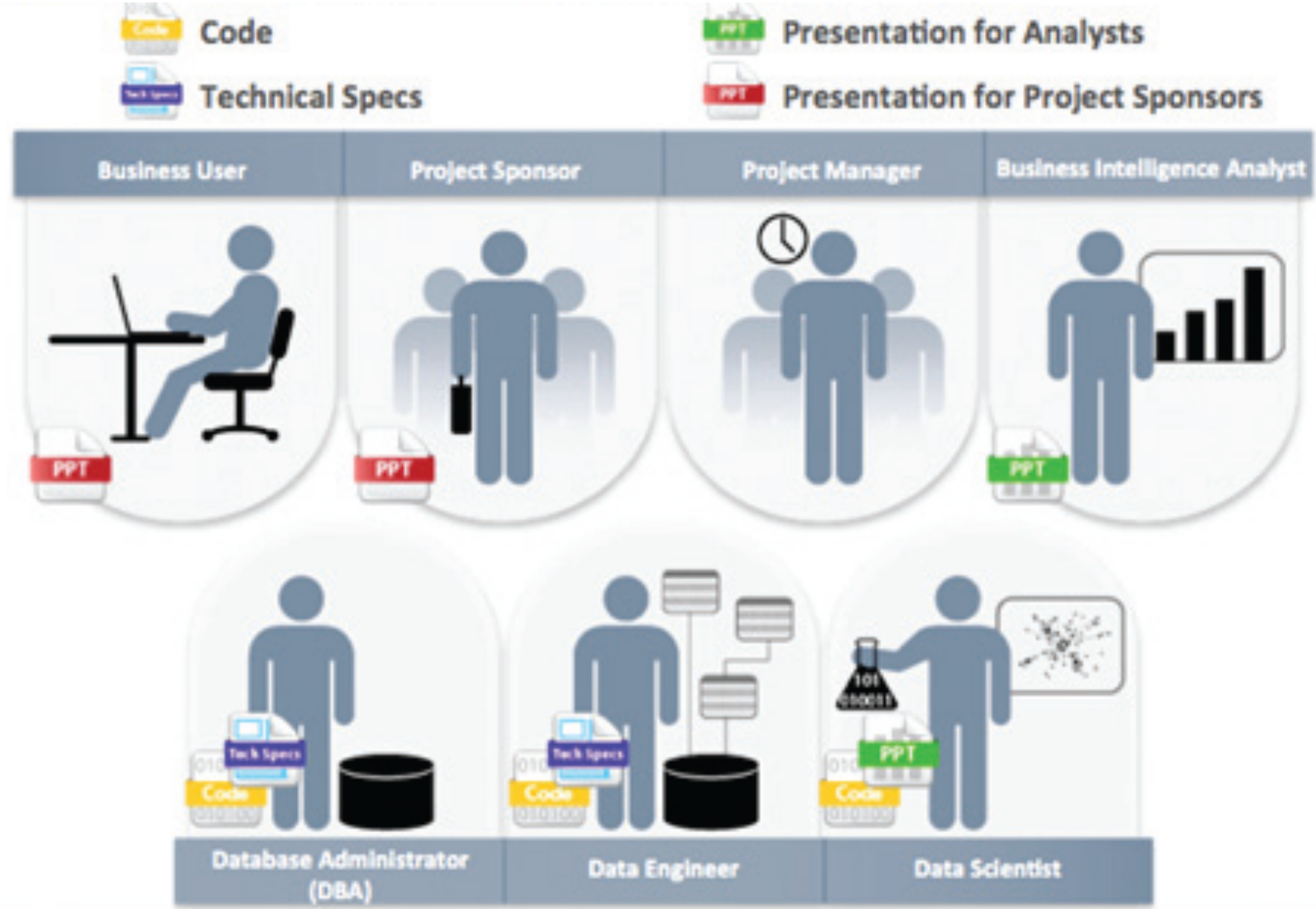
Source: EMC Education Services, Data Science and Big Data Analytics: Discovering, Analyzing, Visualizing and Presenting Data, Wiley, 2015

Overview of Data Analytics Lifecycle

1. Discovery
2. Data preparation
3. Model planning
4. Model building
5. Communicate results
6. Operationalize

Source: EMC Education Services, Data Science and Big Data Analytics: Discovering, Analyzing, Visualizing and Presenting Data, Wiley, 2015

Key Outputs from a Successful Analytics Project



Source: EMC Education Services, Data Science and Big Data Analytics: Discovering, Analyzing, Visualizing and Presenting Data, Wiley, 2015

Data Mining Process

Data Mining Process

- A manifestation of best practices
- A systematic way to conduct DM projects
- Different groups has different versions
- Most common standard processes:
  - CRISP-DM
    (Cross-Industry Standard Process for Data Mining)
  - SEMMA
    (Sample, Explore, Modify, Model, and Assess)
  - KDD
    (Knowledge Discovery in Databases)

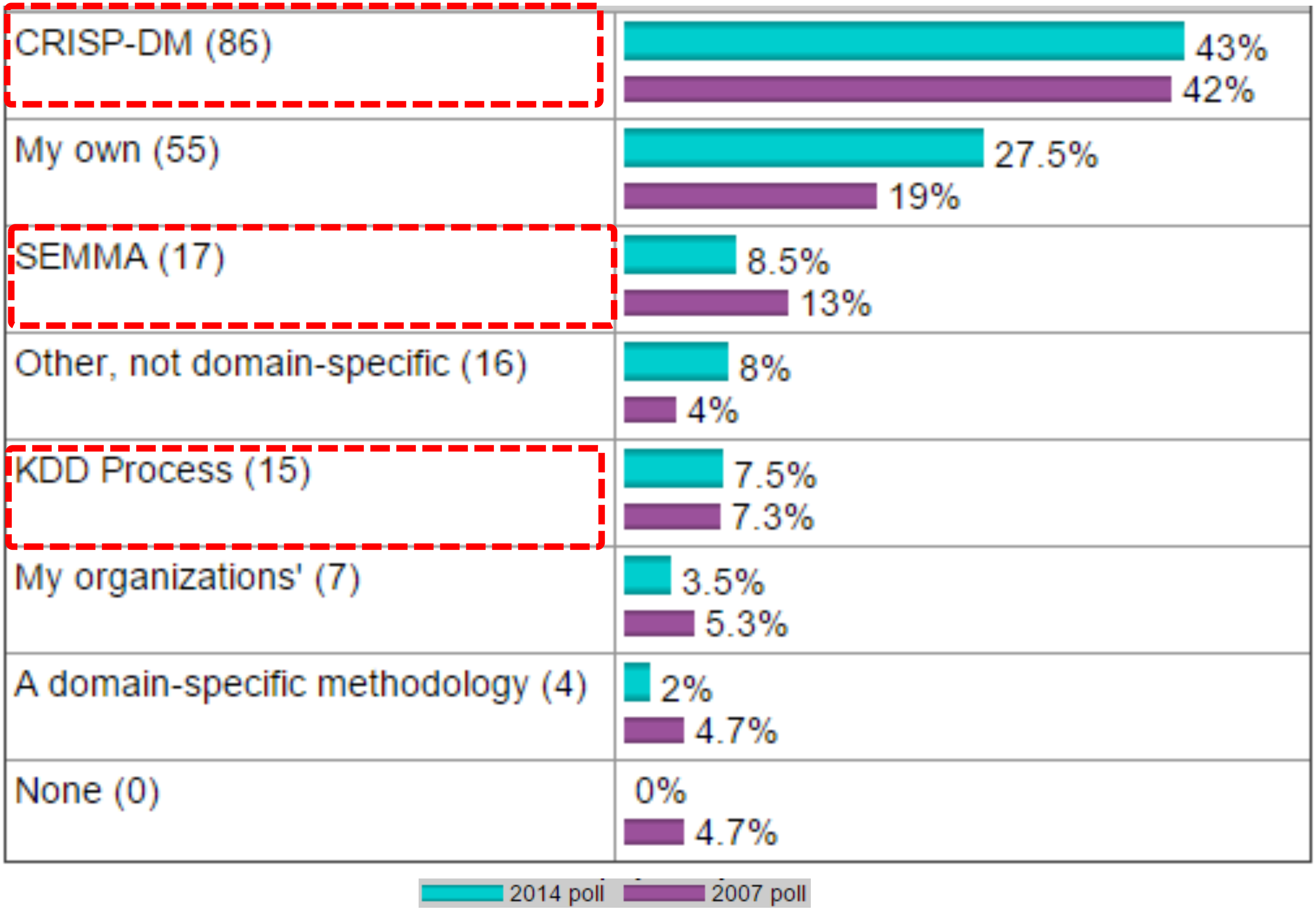
Source: Turban et al. (2011), Decision Support and Business Intelligence Systems

Data Mining Process (SOP of DM)

What main methodology are you using for your **analytics,**
data mining,
or **data science** projects ?

Source: http://www.kdnuggets.com/polls/2014/analytics-data-mining-data-science-methodology.html

Data Mining Process



Source: http://www.kdnuggets.com/polls/2014/analytics-data-mining-data-science-methodology.html

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.

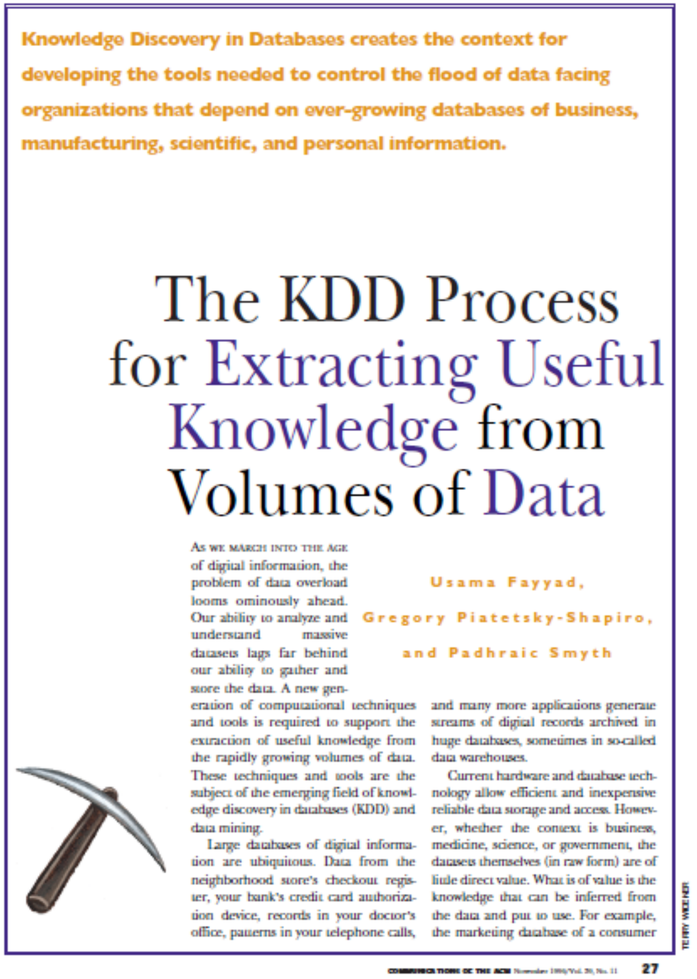
Fayyad, U., Piatetsky-Shapiro, G., & Smyth, P. (1996).

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

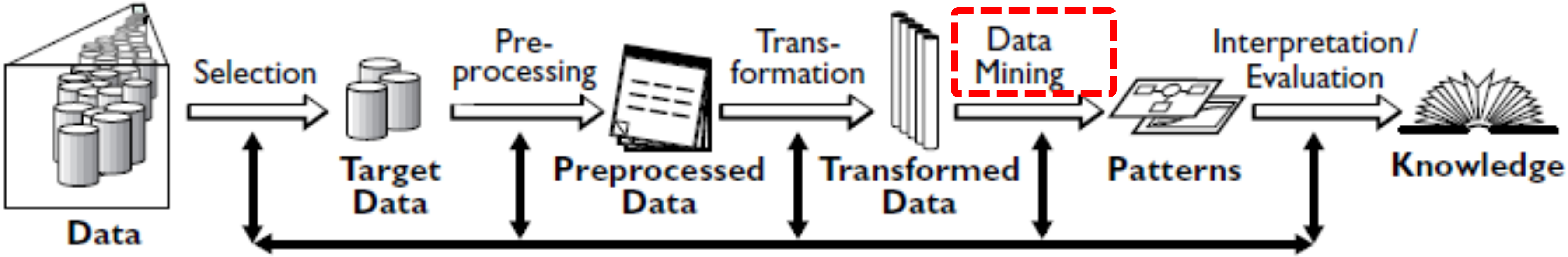
TERRY WIDENER

COMMUNICATIONS OF THE ACM November 1996/Vol. 39, No. 11 27

Data Mining

Knowledge Discovery in Databases (KDD) Process

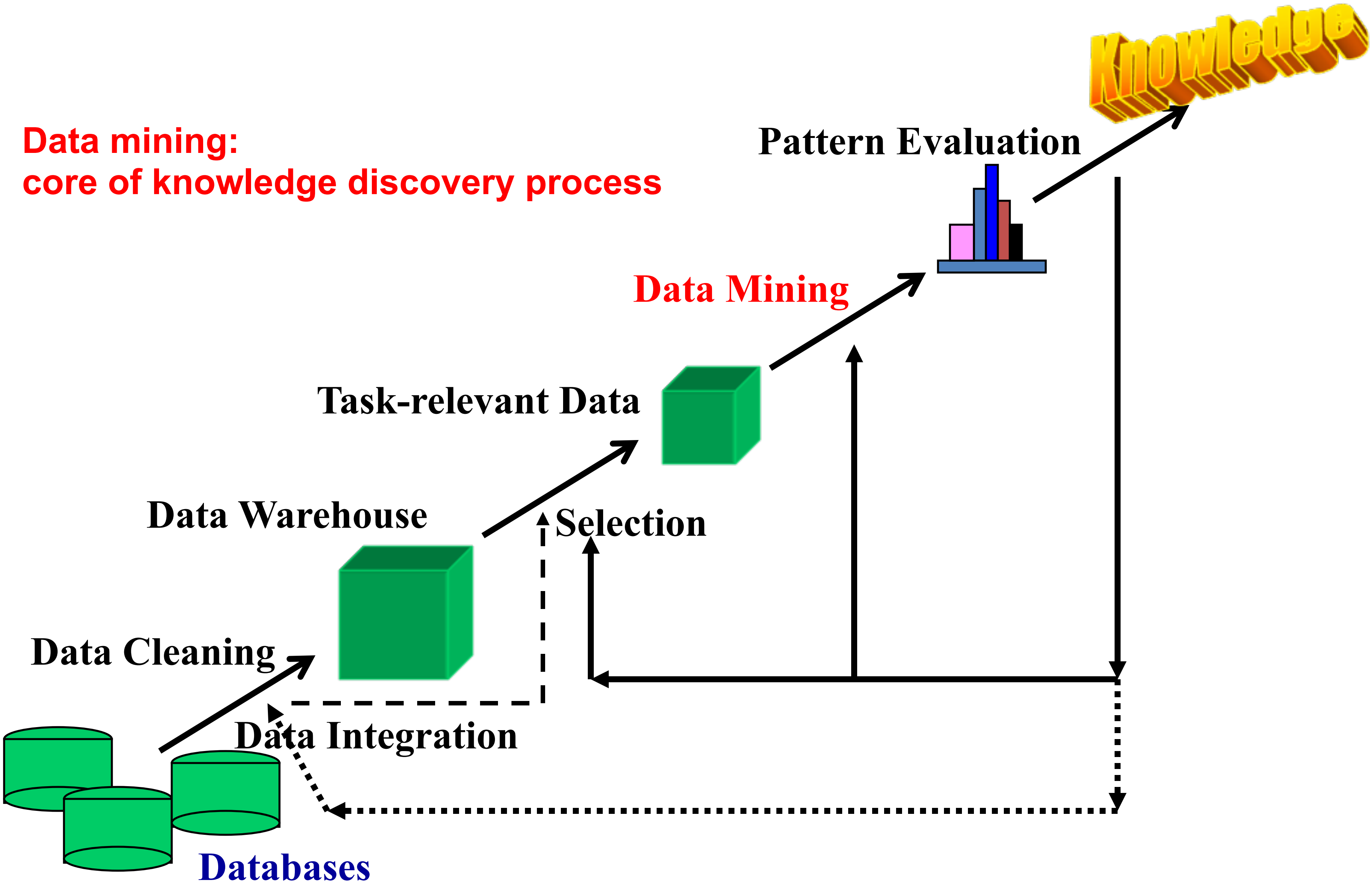
(Fayyad et al., 1996)



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.

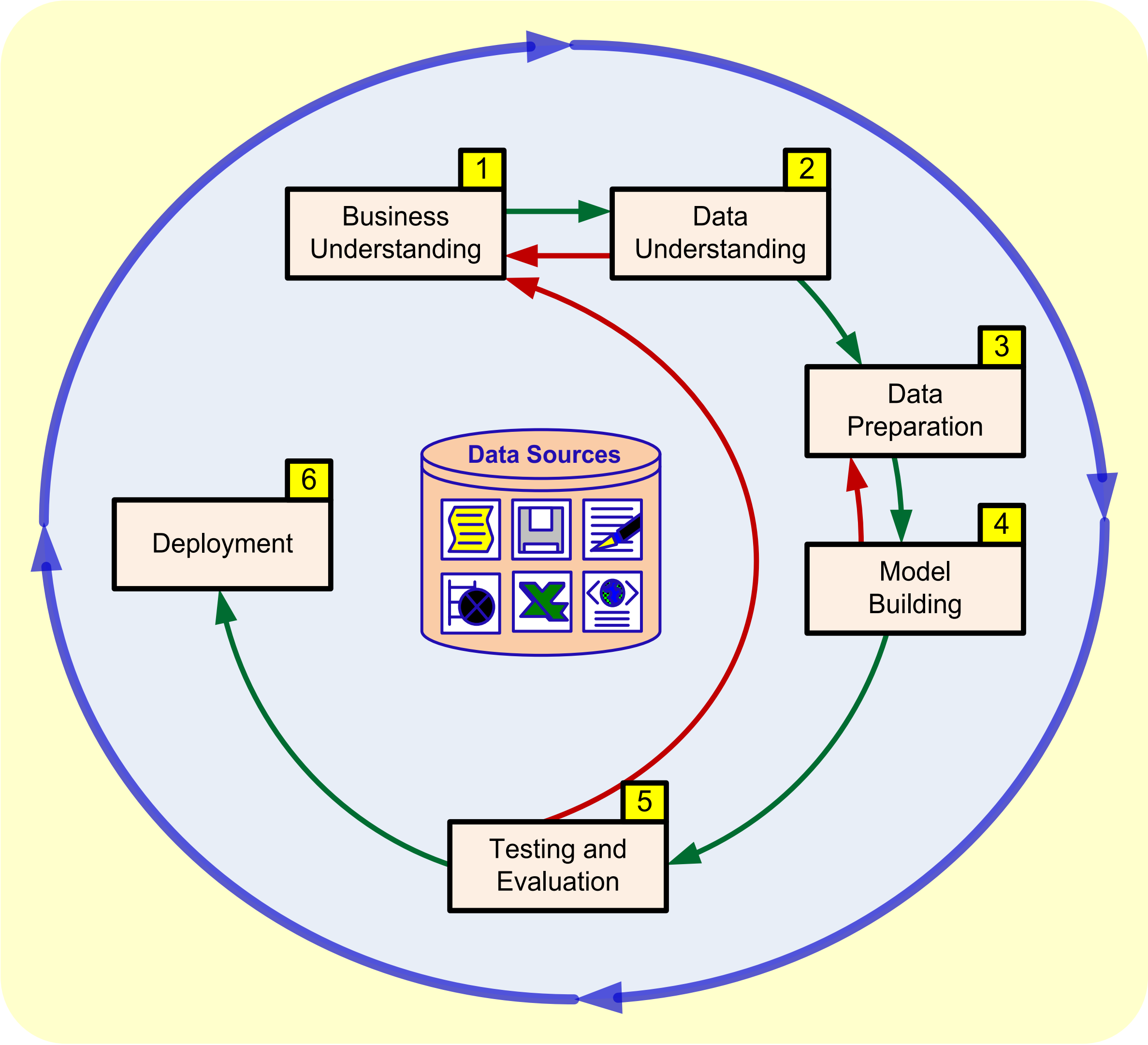
Knowledge Discovery (KDD) Process

Data mining:
core of knowledge discovery process



Source: Han & Kamber (2006)

Data Mining Process: CRISP-DM



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

Data Mining Process: CRISP-DM

Step 1: Business Understanding

Step 2: Data Understanding

Step 3: Data Preparation (!)

Accounts for ~85% of total project time

Step 4: Model Building

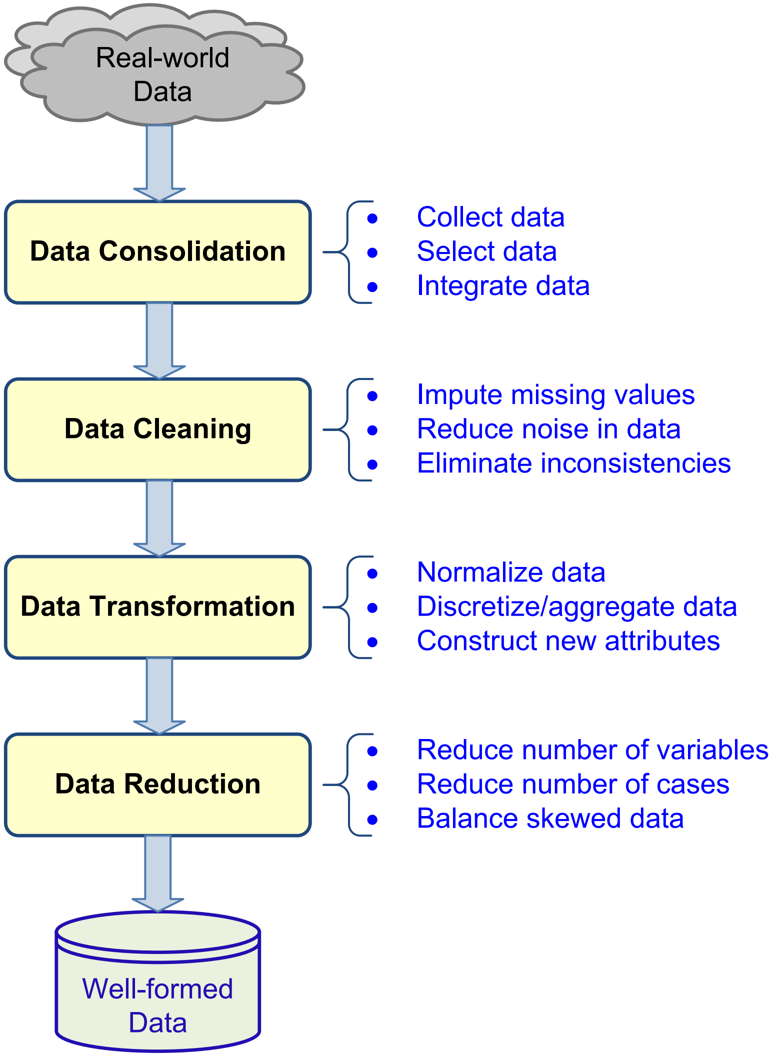
Step 5: Testing and Evaluation

Step 6: Deployment

- The process is highly repetitive and experimental (DM: art versus science?)

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

Data Preparation – A Critical DM Task



- Collect data
- Select data
- Integrate data

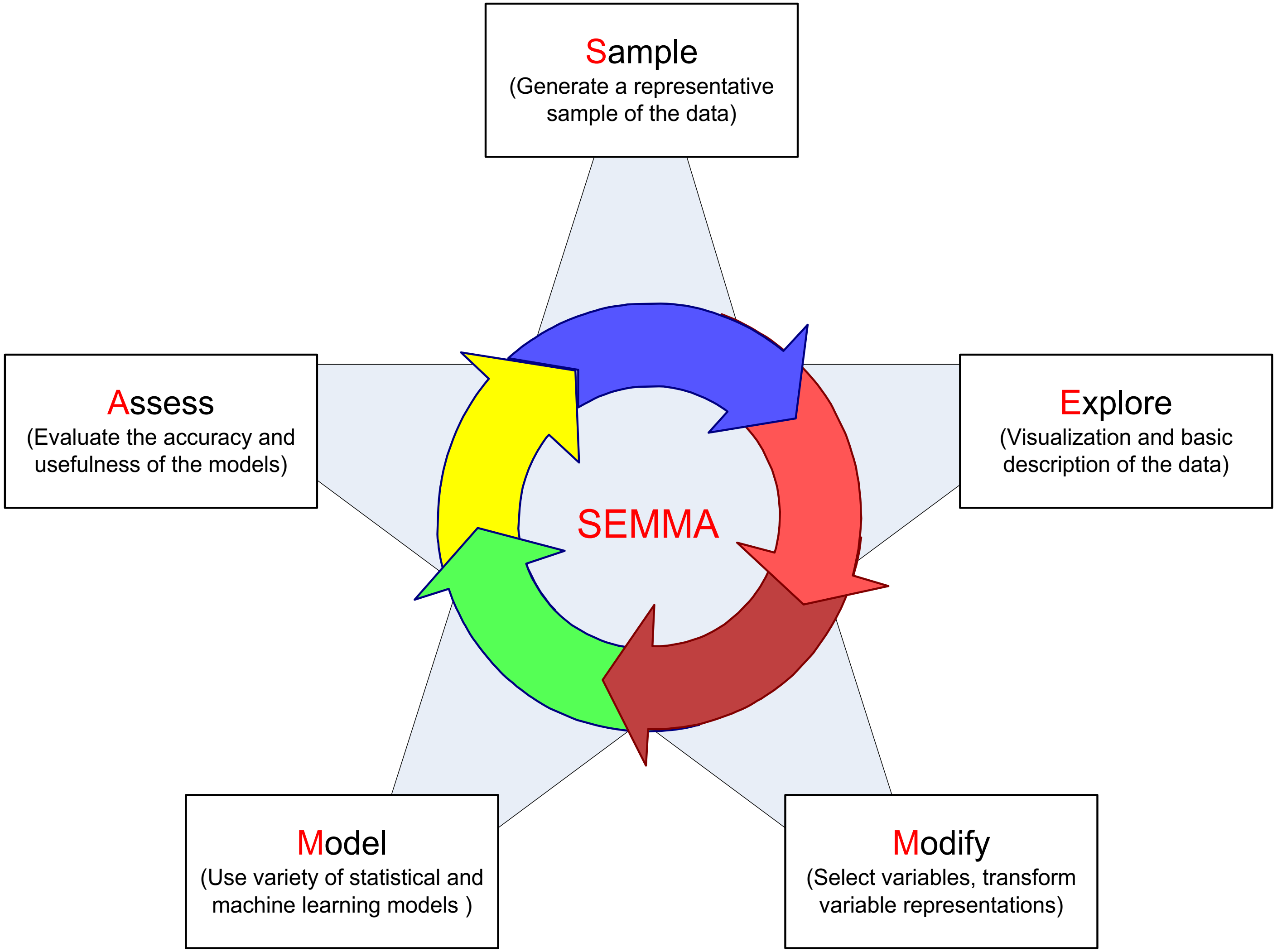
- Impute missing values
- Reduce noise in data
- Eliminate inconsistencies

- Normalize data
- Discretize/aggregate data
- Construct new attributes

- Reduce number of variables
- Reduce number of cases
- Balance skewed data

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

Data Mining Process: SEMMA



Sample
(Generate a representative sample of the data)

Explore
(Visualization and basic description of the data)

Modify
(Select variables, transform variable representations)

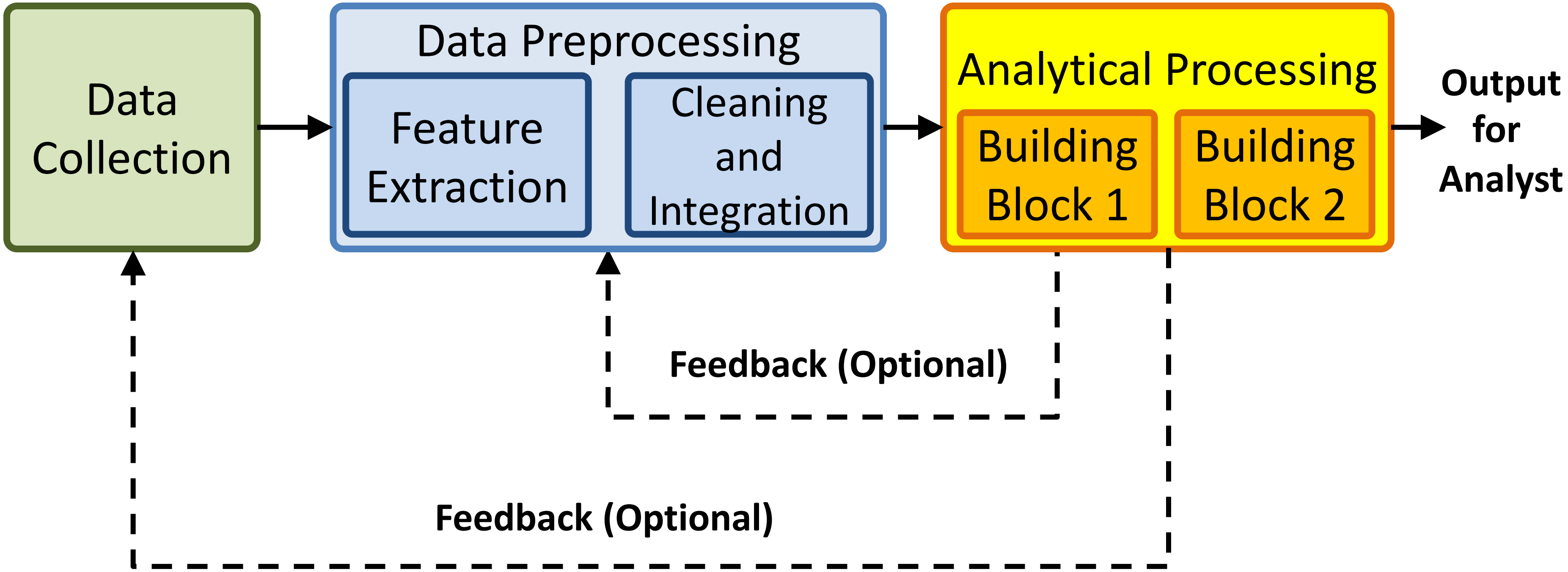
Model
(Use variety of statistical and machine learning models)

Assess
(Evaluate the accuracy and usefulness of the models)

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

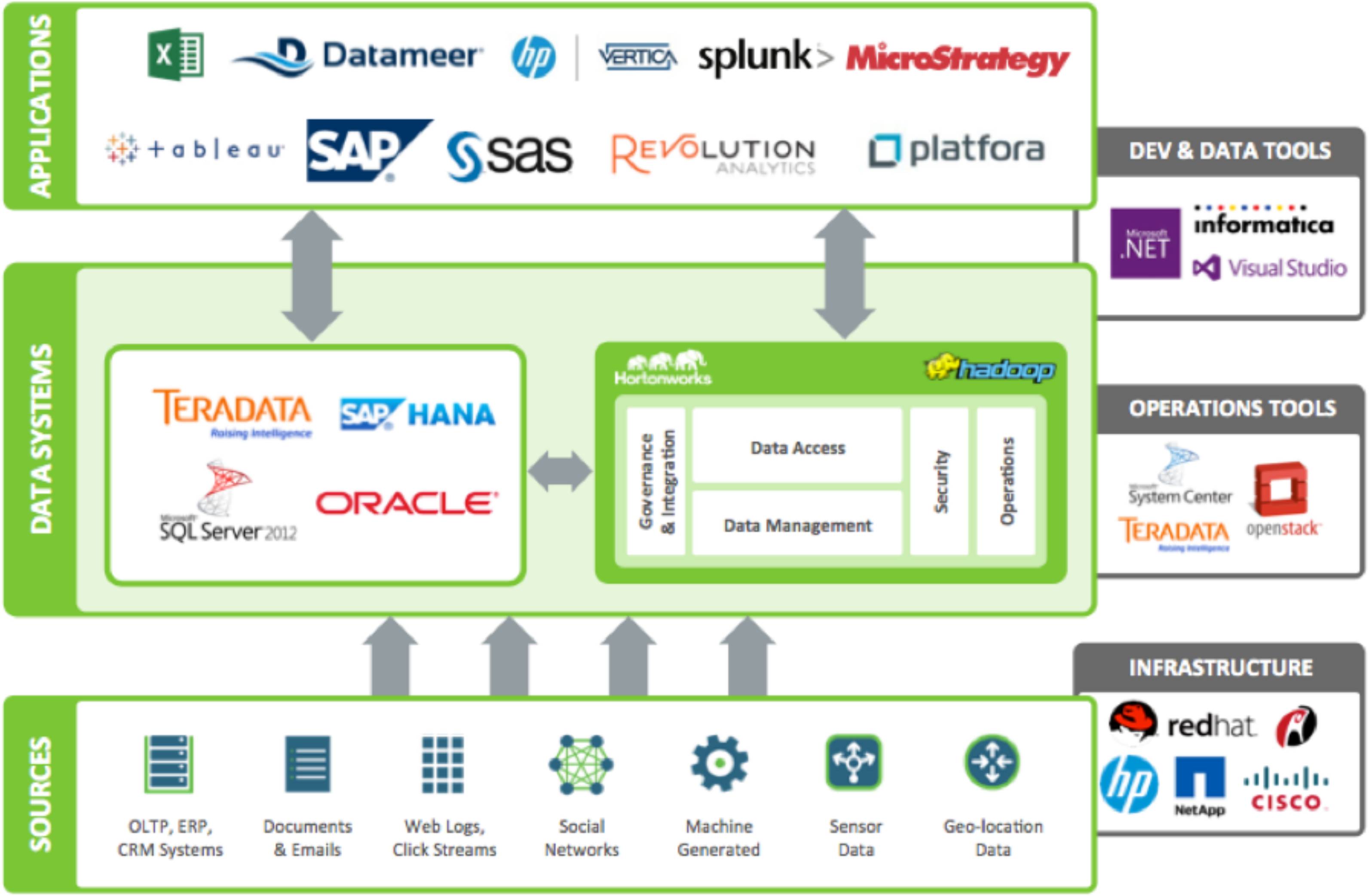
Data Mining Processing Pipeline

(Charu Aggarwal, 2015)



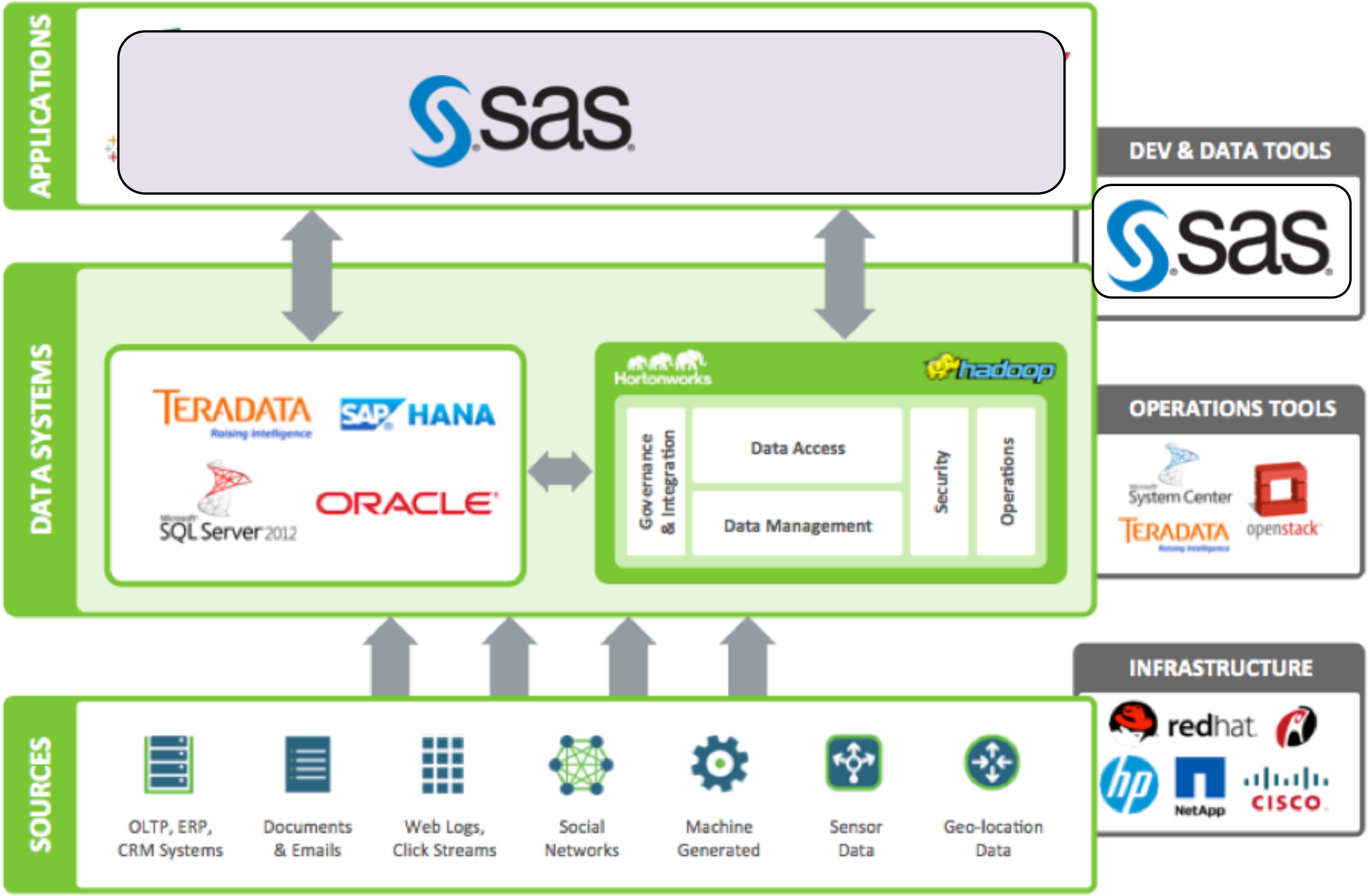
Source: Charu Aggarwal (2015), Data Mining: The Textbook Hardcover, Springer

SAS Big data Strategy – SAS areas



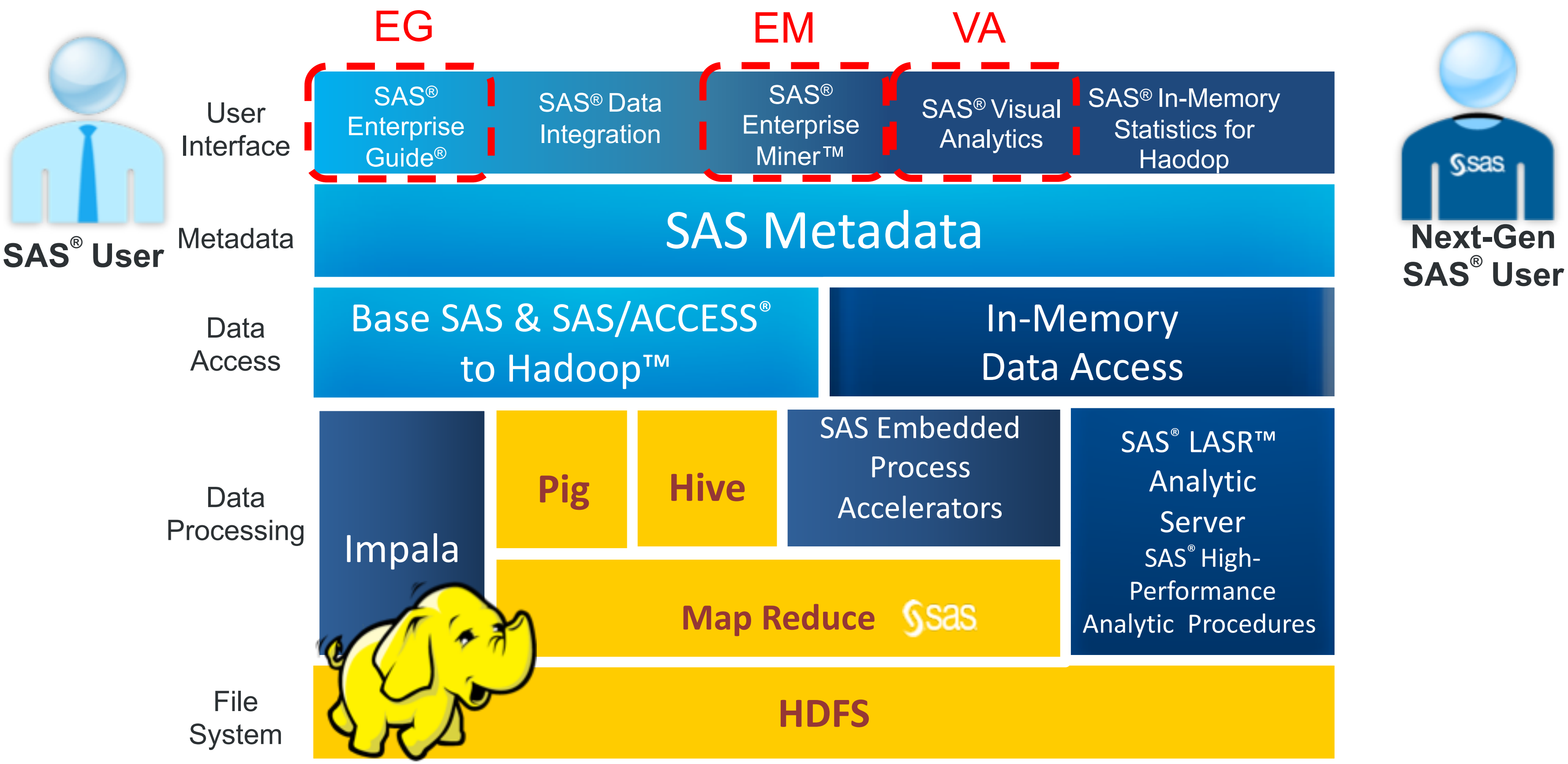
Source: Deepak Ramanathan (2014), SAS Modernization architectures - Big Data Analytics

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

Summary

- AI
- Big Data Analytics

References

- Ramesh Sharda, Dursun Delen, and Efraim Turban (2017), Business Intelligence, Analytics, and Data Science: A Managerial Perspective, 4th Edition, Pearson.
- Jared Dean (2014), Big Data, Data Mining, and Machine Learning: Value Creation for Business Leaders and Practitioners, Wiley.
- Mehmet Kaya, Jalal Kawash, Suheil Khoury, and Min-Yuh Day (2018), Social Network Based Big Data Analysis and Applications, Lecture Notes in Social Networks, Springer International Publishing.
- Varun Grover, Roger HL Chiang, Ting-Peng Liang, and Dongsong Zhang (2018), "Creating Strategic Business Value from Big Data Analytics: A Research Framework", Journal of Management Information Systems, 35, no. 2, pp. 388-423.
- Ting-Peng Liang and Yu-Hsi Liu (2018), "Research Landscape of Business Intelligence and Big Data analytics: A bibliometrics study", Expert Systems with Applications, 111, no. 30, pp. 2-10.
- Stuart Russell and Peter Norvig (2016) , Artificial Intelligence: A Modern Approach, 3rd Edition, Pearson International.
- Javier Mata, Ignacio de Miguel, Ramón J. Durán, Noemí Merayo, Sandeep Kumar Singh, Admela Jukan, and Mohit Chamania (2018), "Artificial intelligence (AI) methods in optical networks: A comprehensive survey", Optical Switching and Networking, 28, pp. 43-57
- Stephan Kudyba (2014), Big Data, Mining, and Analytics: Components of Strategic Decision Making, Auerbach Publications.