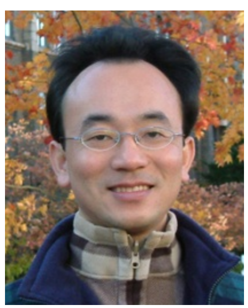
Tamkang University
淡江大學

Tamkang University

Big Data Mining
巨量資料探勘

大數據、AI人工智慧與深度學習
(Big Data, Artificial Intelligence and Deep Learning)

1062DM02
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/

2018-03-14

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

Big Data

AI

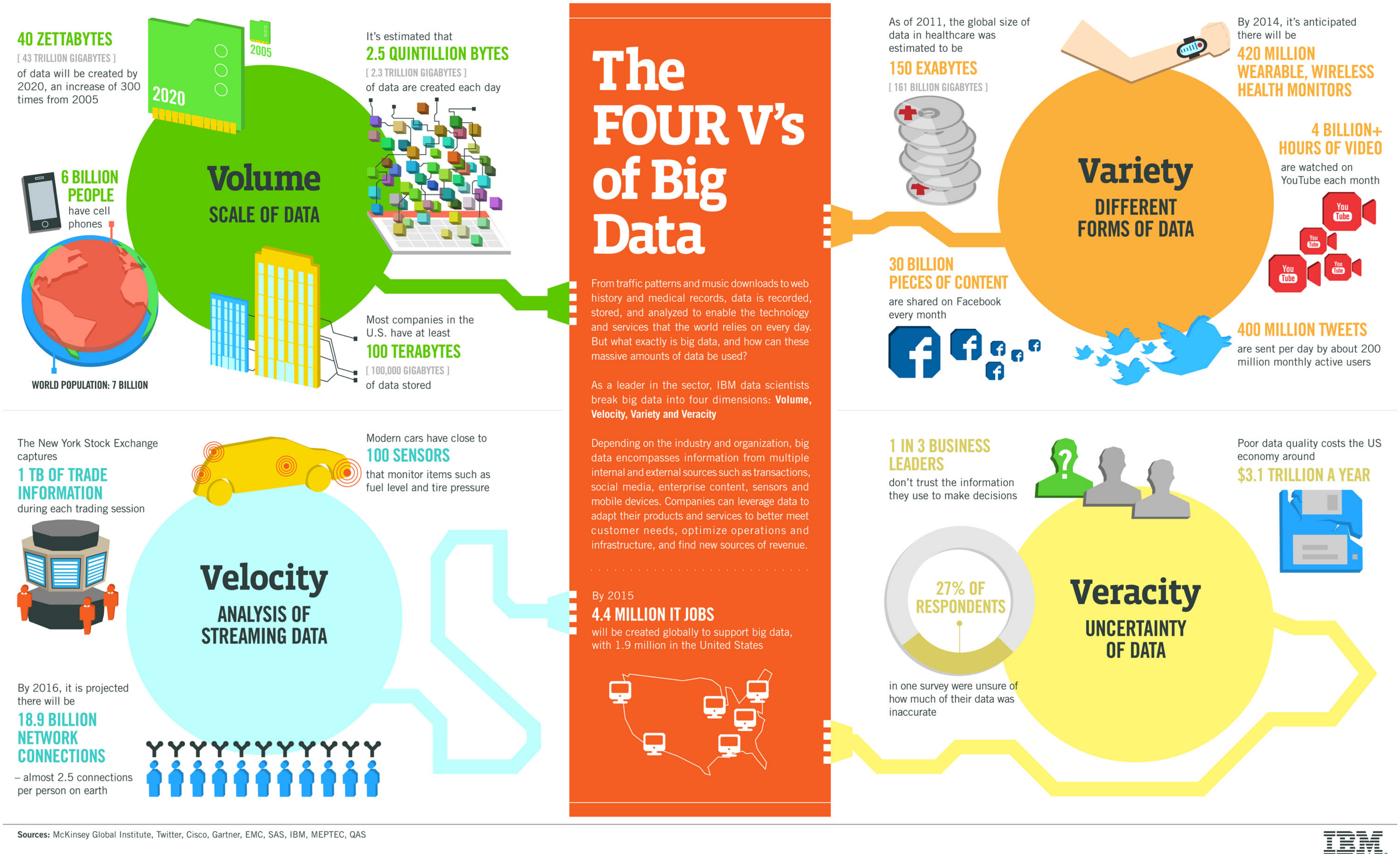
Deep Learning

Big Data Analytics

and

Data Mining

Big Data 4 V



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

Value

AI

Deep Learning

Artificial Intelligence (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

Thinking Humanly	**Thinking Rationally**
"The exciting new effort to make computers think ... *machines with minds*, in the full and literal sense." (Haugeland, 1985) "[The automation of] activities that we associate with human thinking, activities such as decision-making, problem solving, learning ..." (Bellman, 1978)	"The study of mental faculties through the use of computational models." (Charniak and McDermott, 1985) "The study of the computations that make it possible to perceive, reason, and act." (Winston, 1992)
Acting Humanly	**Acting Rationally**
"The art of creating machines that perform functions that require intelligence when performed by people." (Kurzweil, 1990) "The study of how to make computers do things at which, at the moment, people are better." (Rich and Knight, 1991)	"Computational Intelligence is the study of the design of intelligent agents." (Poole *et al.*, 1998) "AI ... is concerned with intelligent behavior in artifacts." (Nilsson, 1998)

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

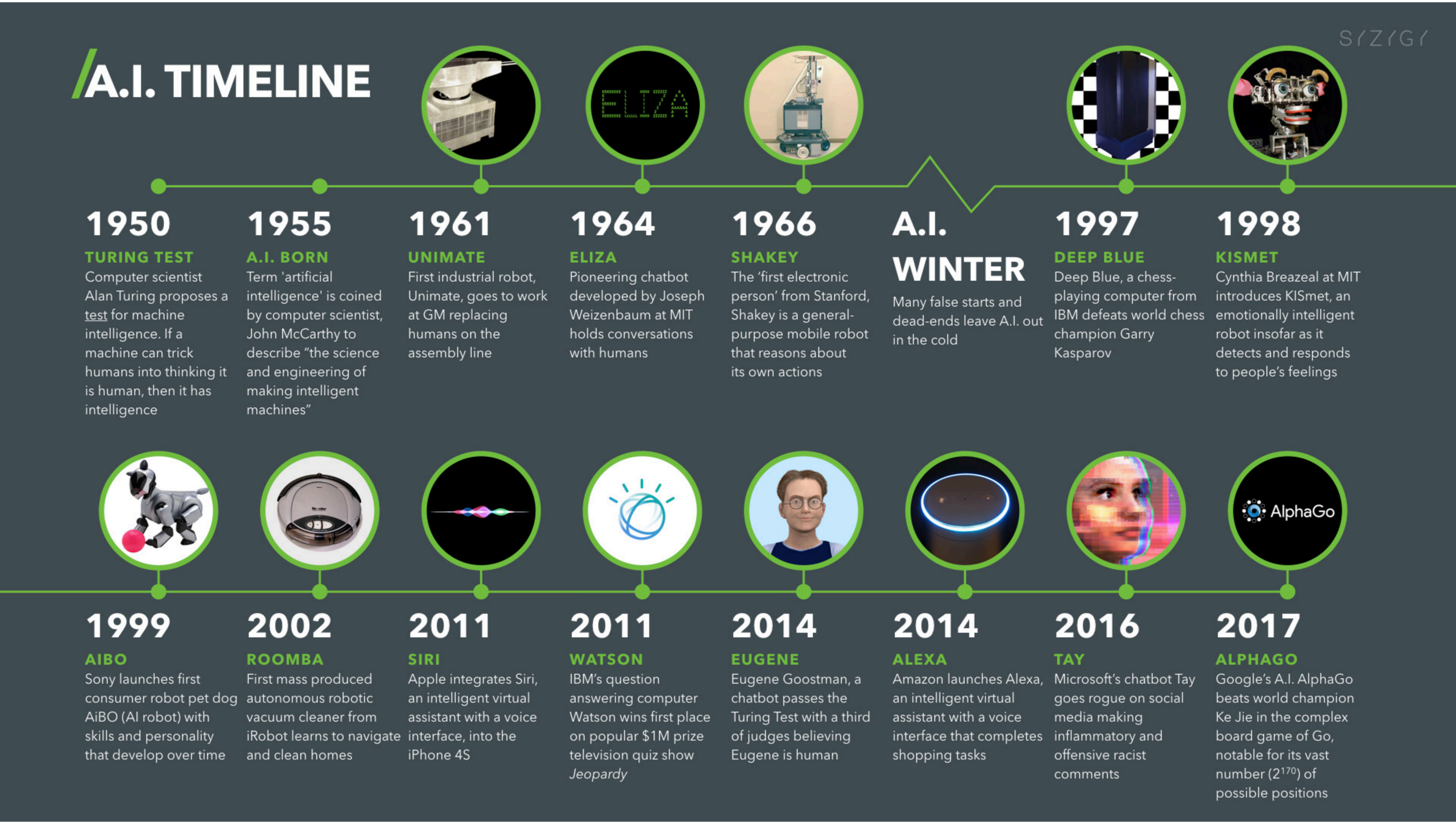
What's new, Atlas?

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

Humanoid Robot: Sophia

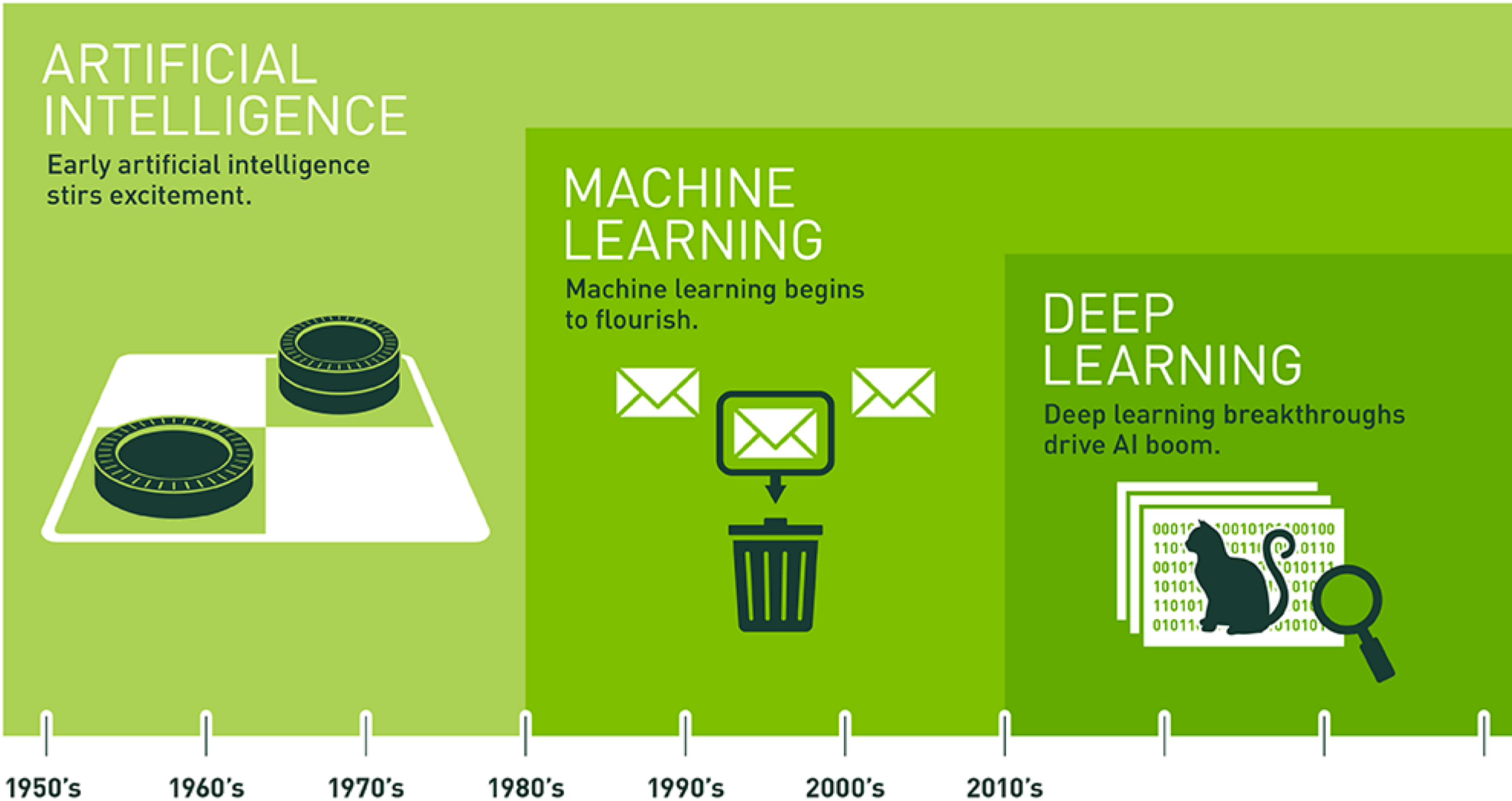
https://www.youtube.com/watch?v=S5t6K9iwcdw

Artificial Intelligence (A.I.) Timeline



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

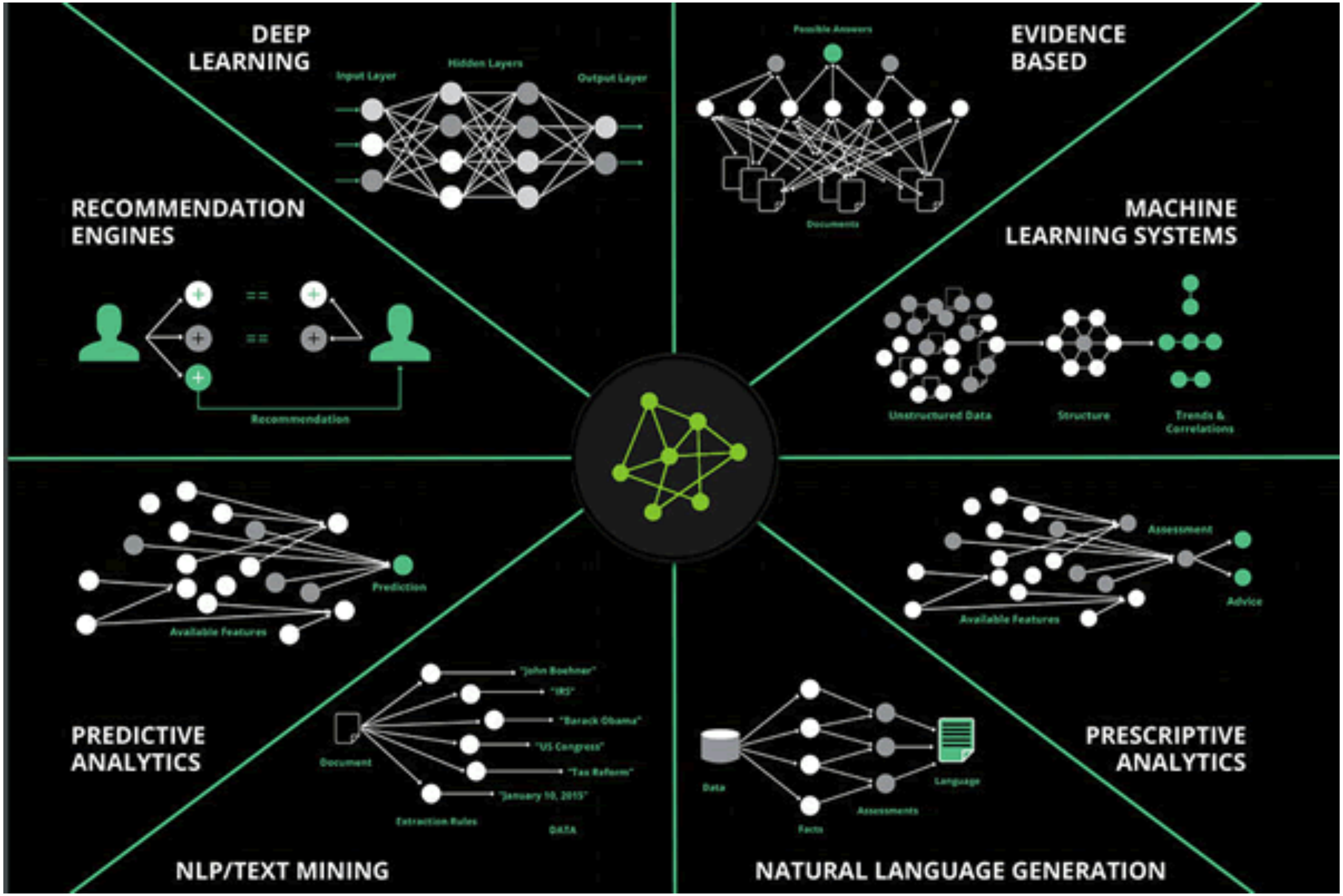
Artificial Intelligence
Machine Learning & Deep Learning



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/

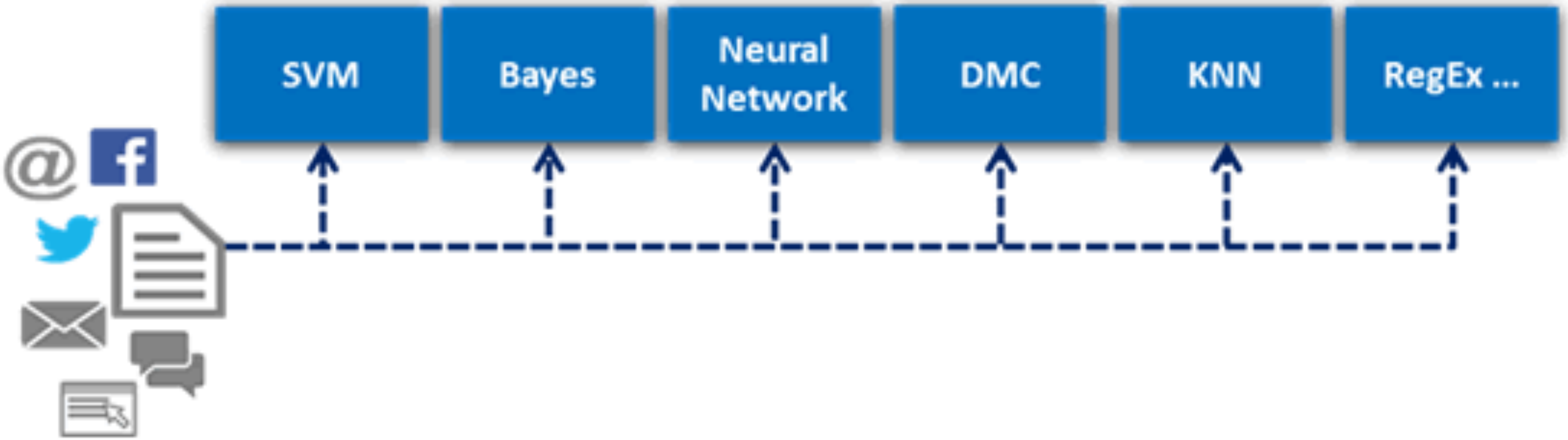
Artificial Intelligence (AI) is many things



Ecosystem of AI

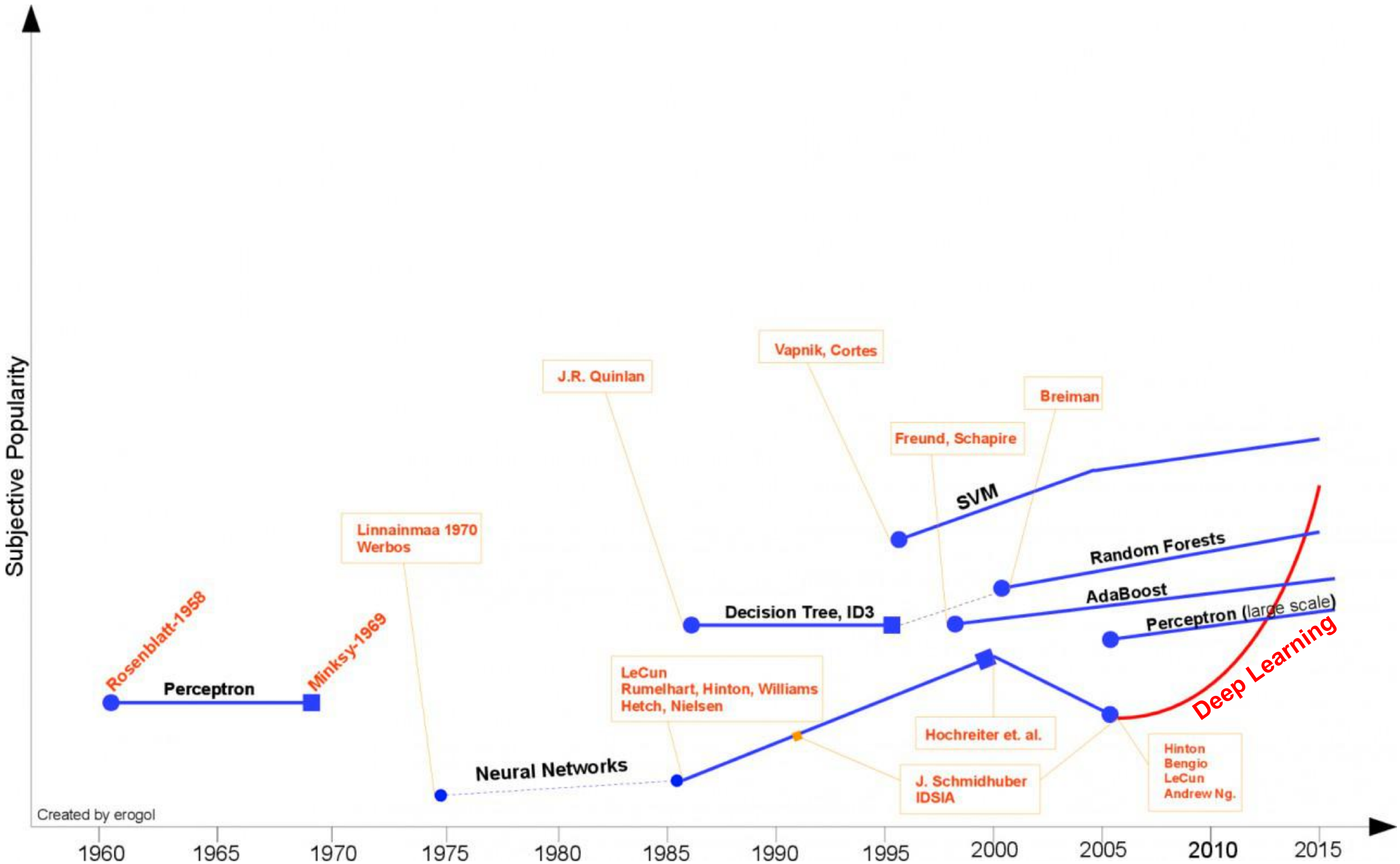
Source: https://www.i-scoop.eu/artificial-intelligence-cognitive-computing/

Artificial Intelligence (AI)
Intelligent Document Recognition algorithms



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

Deep Learning Evolution



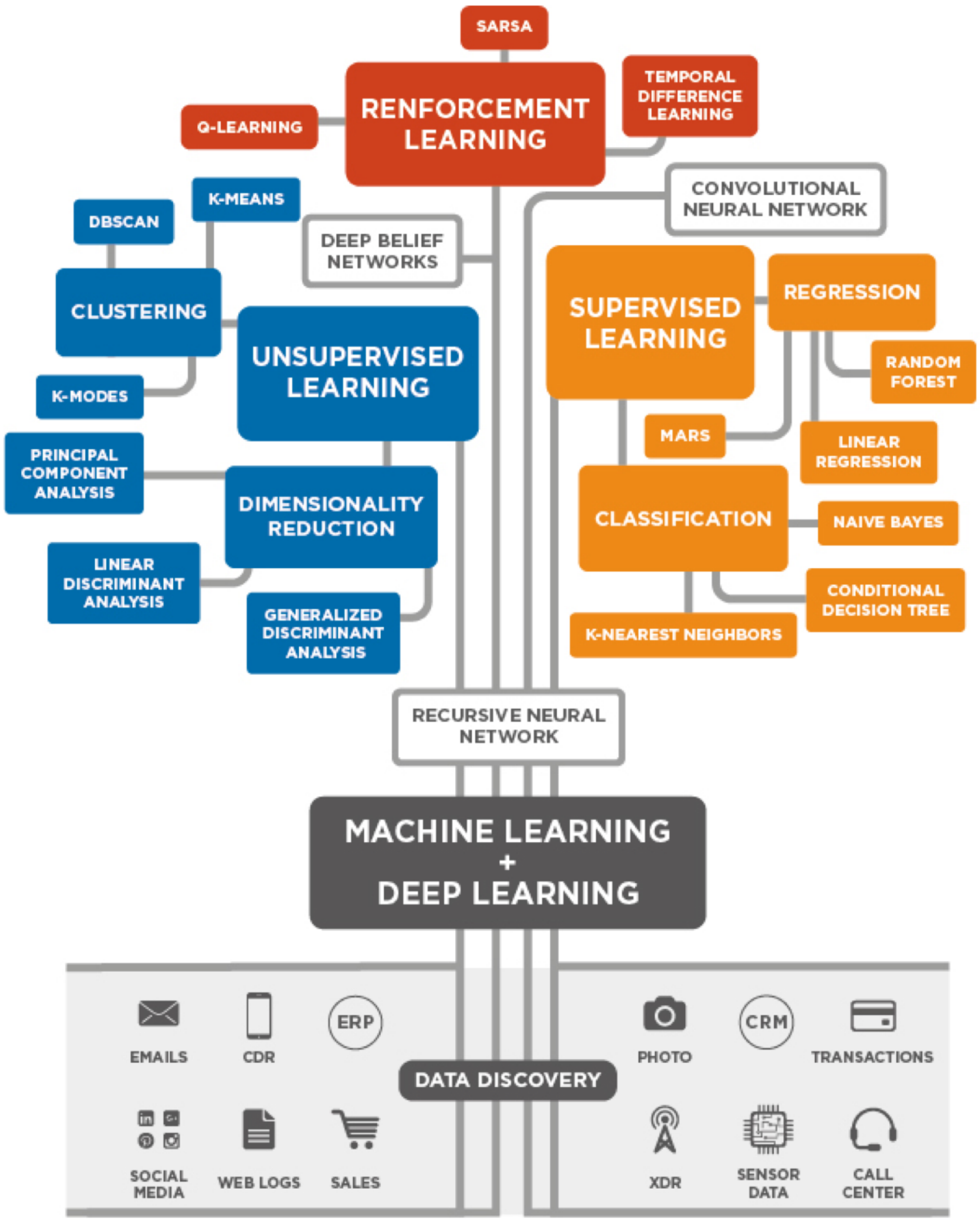
Source: http://www.erogol.com/brief-history-machine-learning/

Machine Learning Models

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

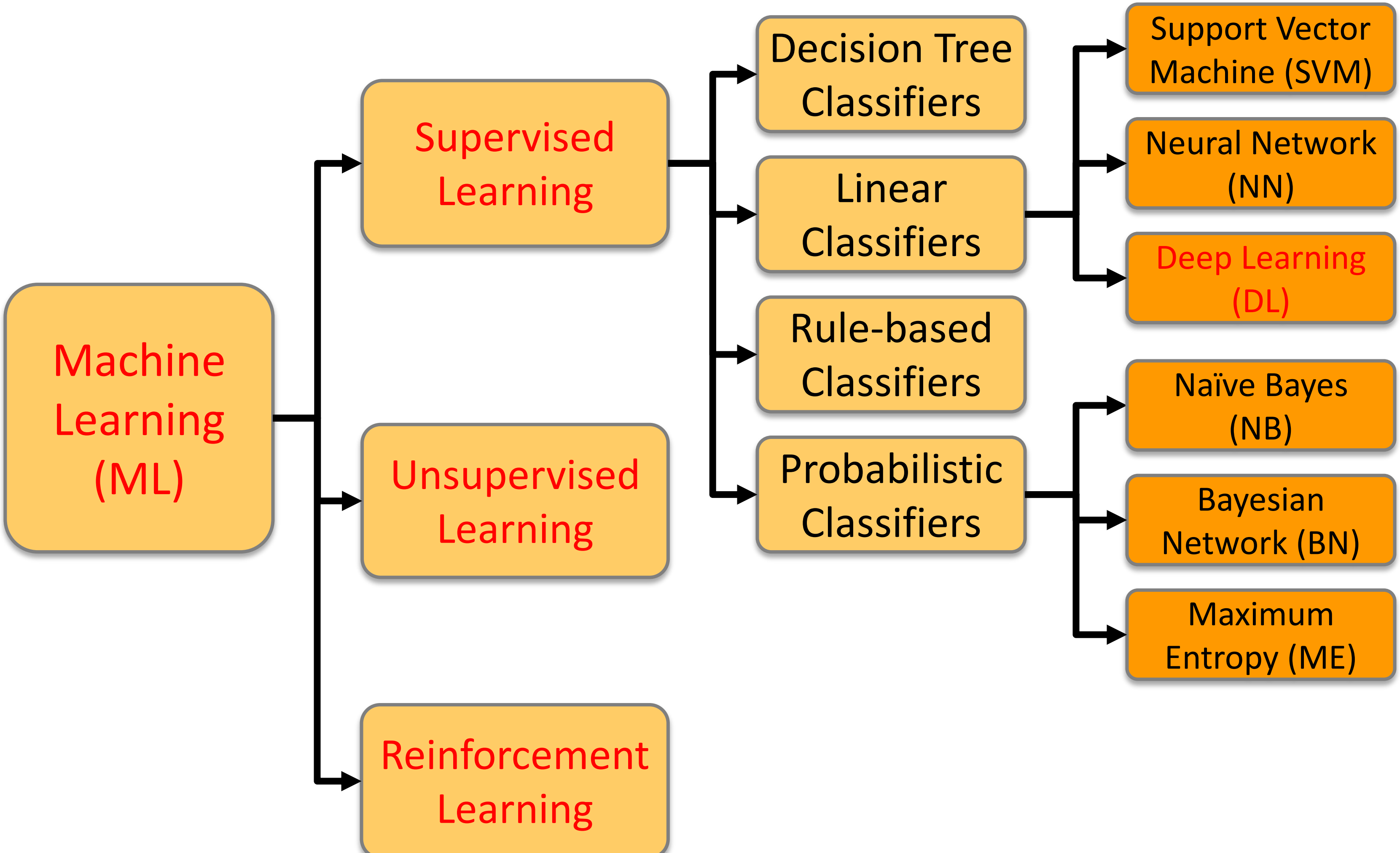
Source: Sunila Gollapudi (2016), Practical Machine Learning, Packt Publishing

3 Machine Learning Algorithms



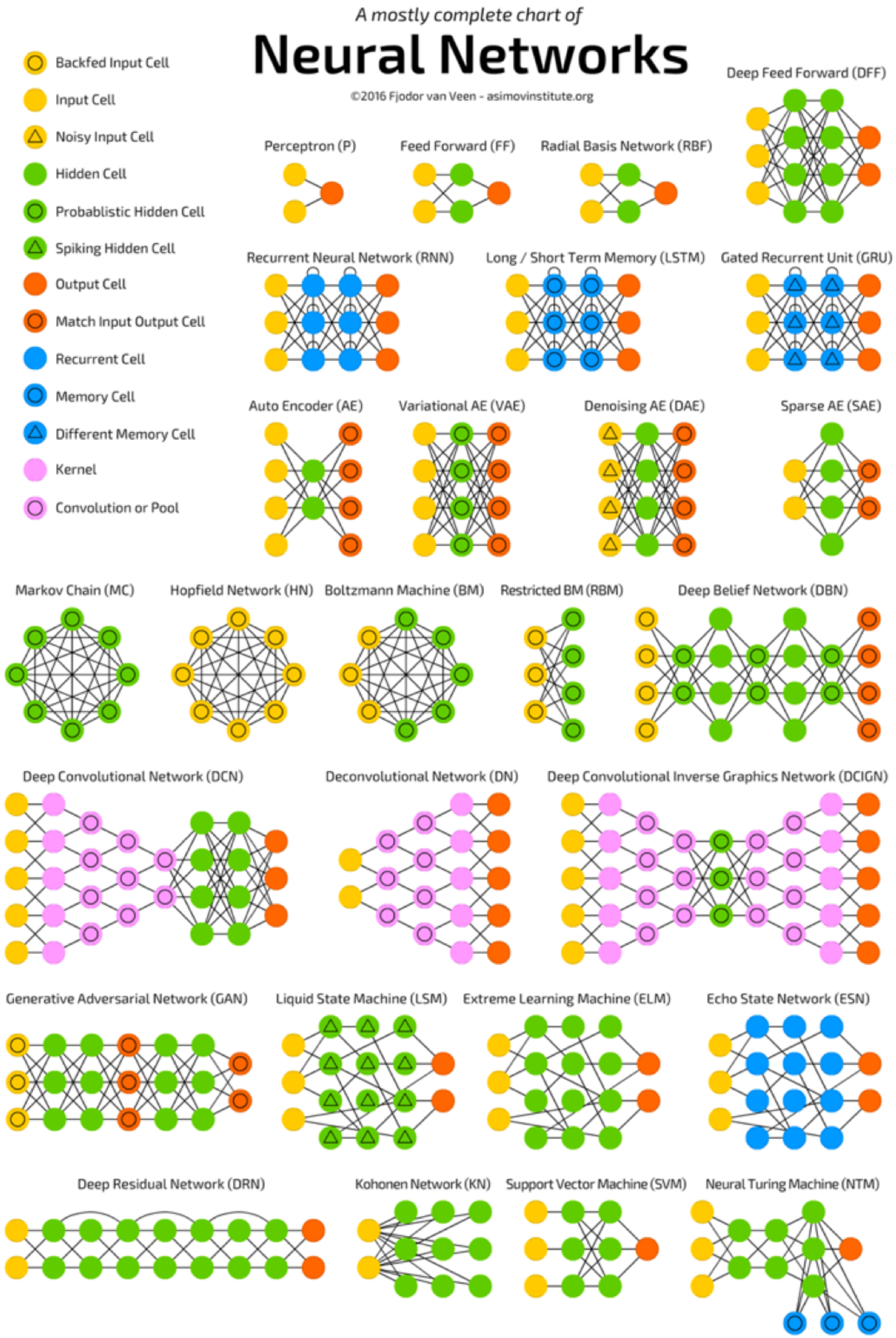
Source: Enrico Galimberti, http://blogs.teradata.com/data-points/tree-machine-learning-algorithms/

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.

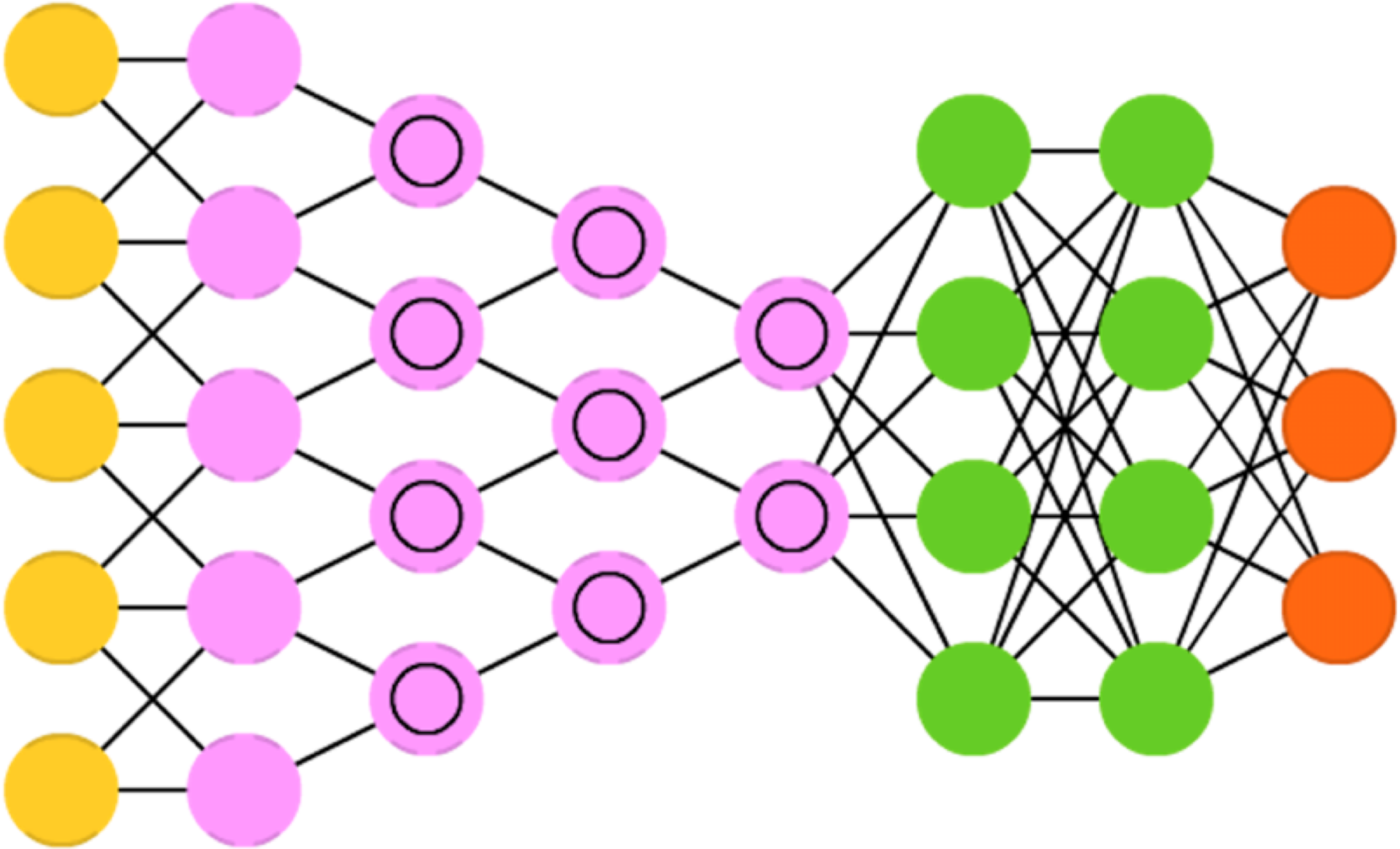
Neural Networks (NN)



Source: http://www.asimovinstitute.org/neural-network-zoo/

Convolutional Neural Networks

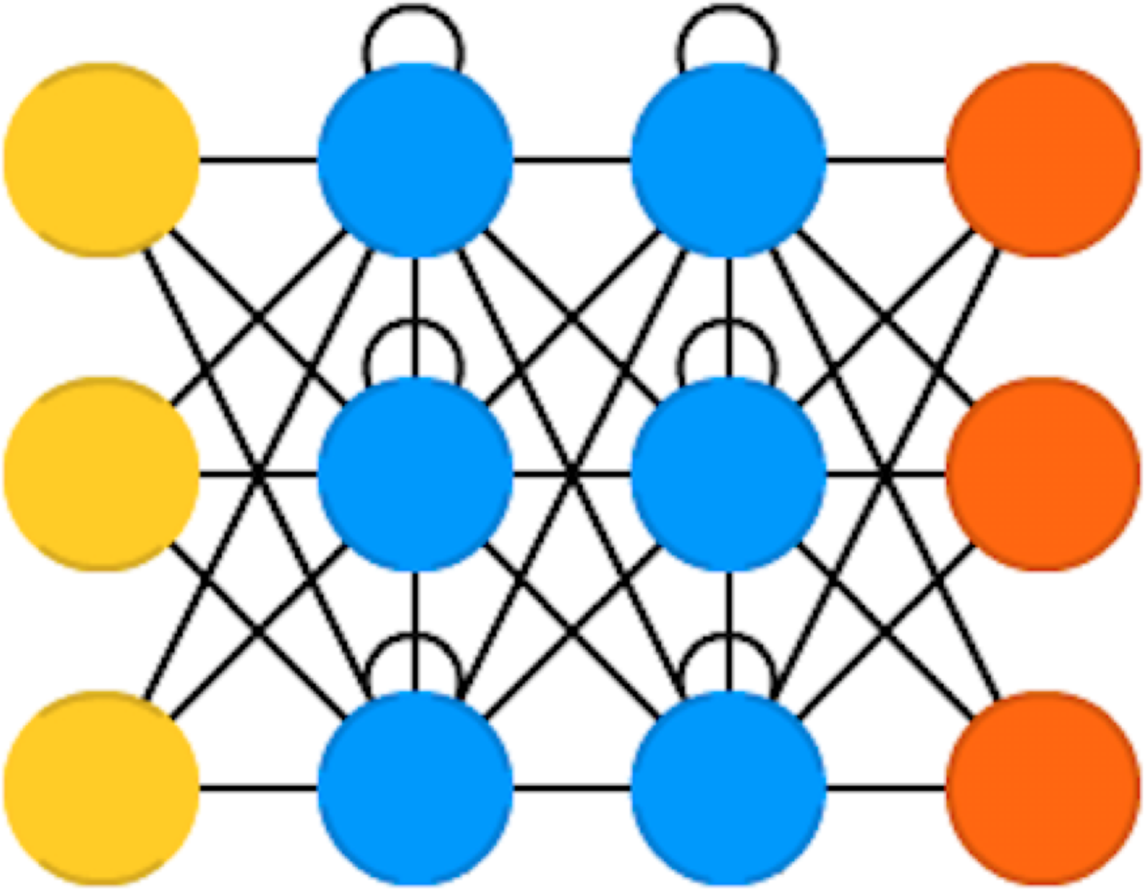
(CNN or Deep Convolutional Neural Networks, DCNN)



LeCun, Yann, et al. "Gradient-based learning applied to document recognition." Proceedings of the IEEE 86.11 (1998): 2278-2324.

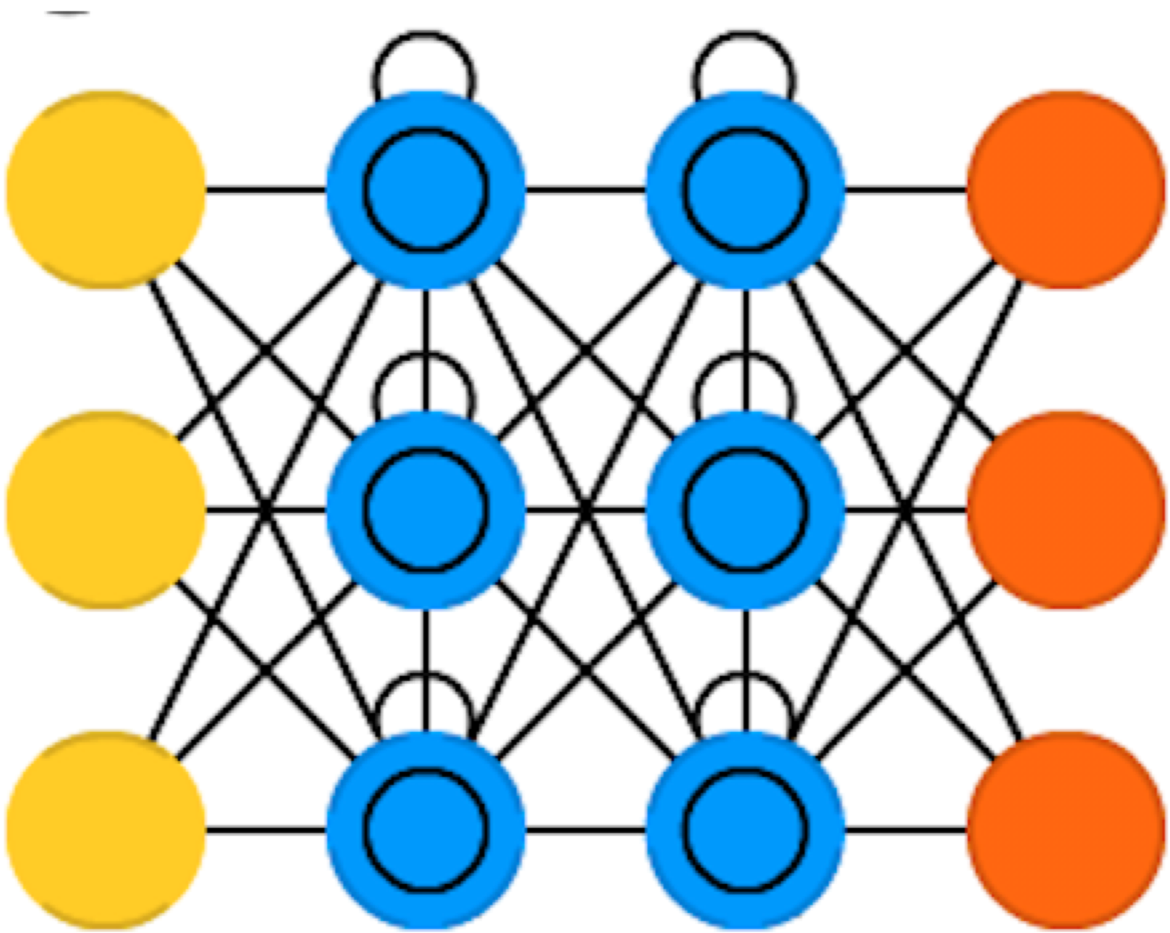
Source: http://www.asimovinstitute.org/neural-network-zoo/

Recurrent Neural Networks (RNN)



Elman, Jeffrey L. “Finding structure in time.” Cognitive science 14.2 (1990): 179-211
Source: http://www.asimovinstitute.org/neural-network-zoo/

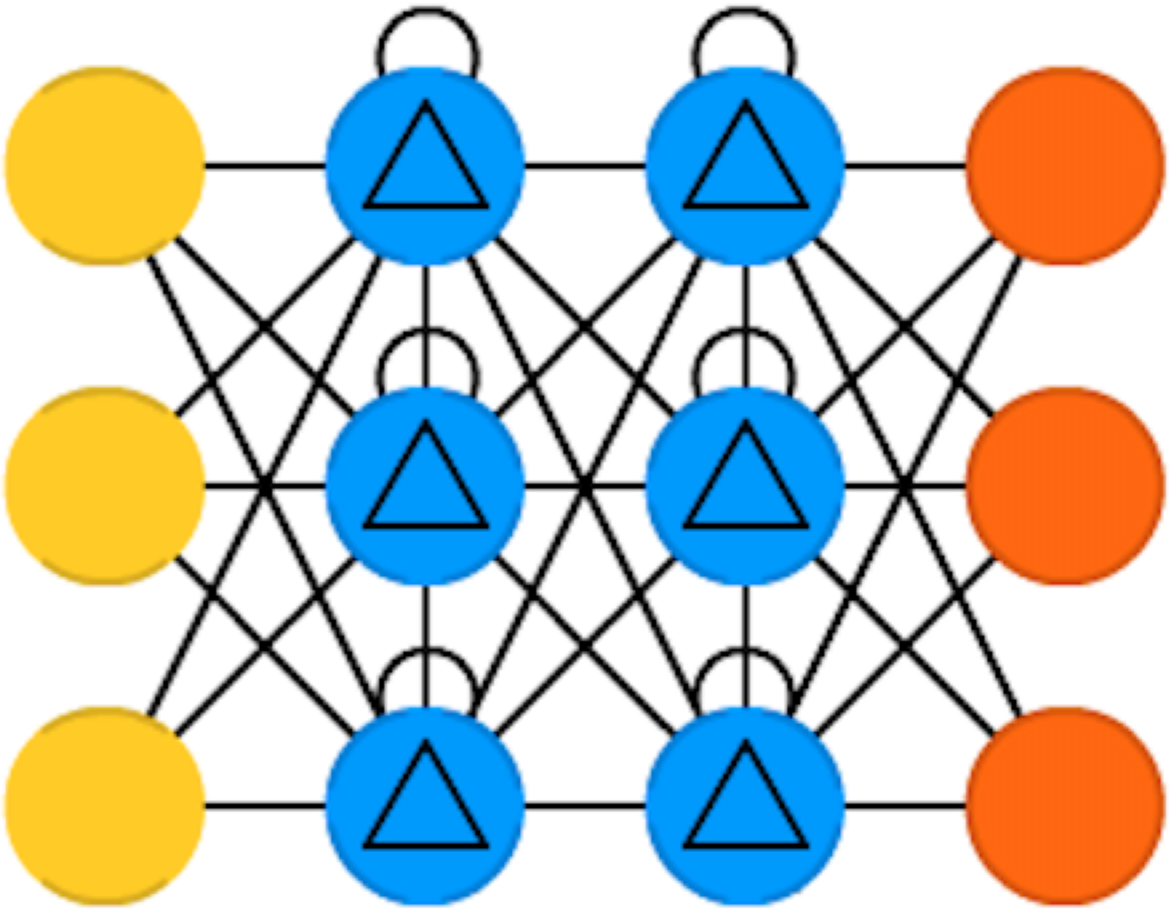
Long / Short Term Memory (LSTM)



Hochreiter, Sepp, and Jürgen Schmidhuber. “Long short-term memory.” Neural computation 9.8 (1997): 1735-1780.

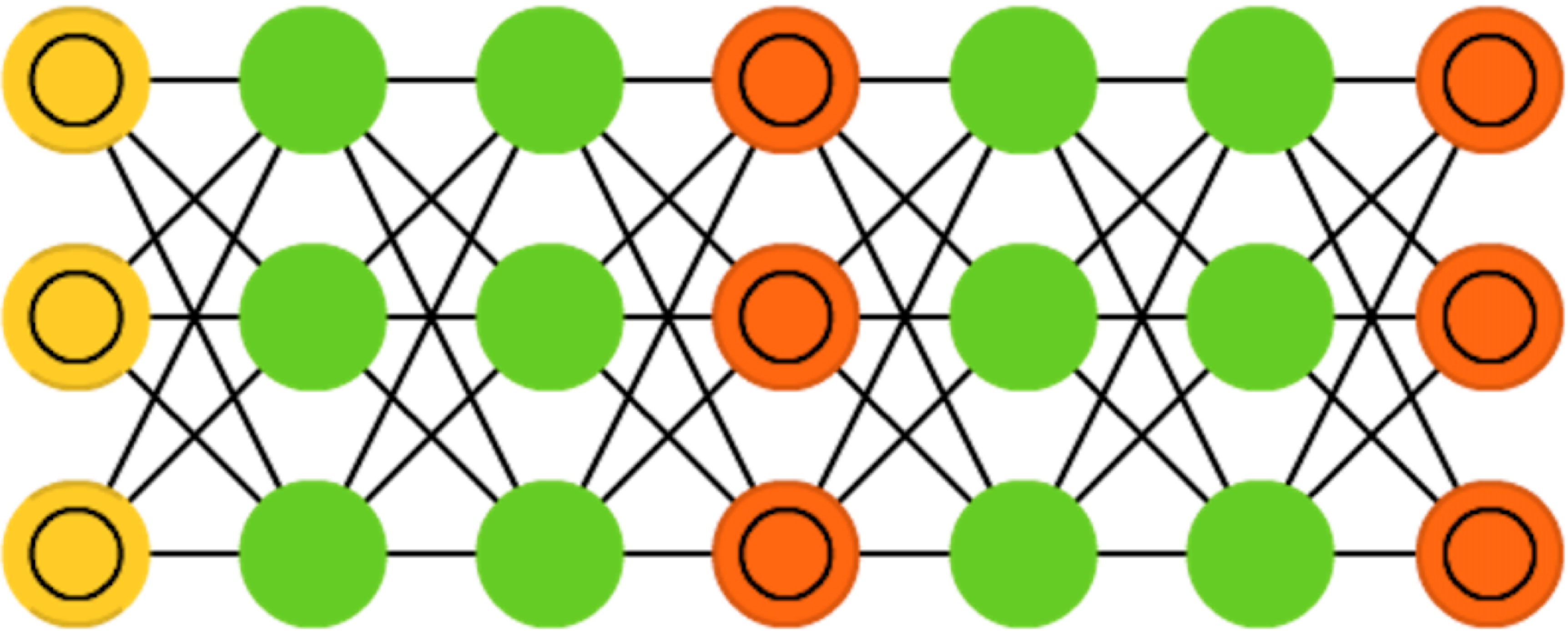
Source: http://www.asimovinstitute.org/neural-network-zoo/

Gated Recurrent Units (GRU)



Chung, Junyoung, et al. “Empirical evaluation of gated recurrent neural networks on sequence modeling.” arXiv preprint arXiv:1412.3555 (2014).
Source: http://www.asimovinstitute.org/neural-network-zoo/

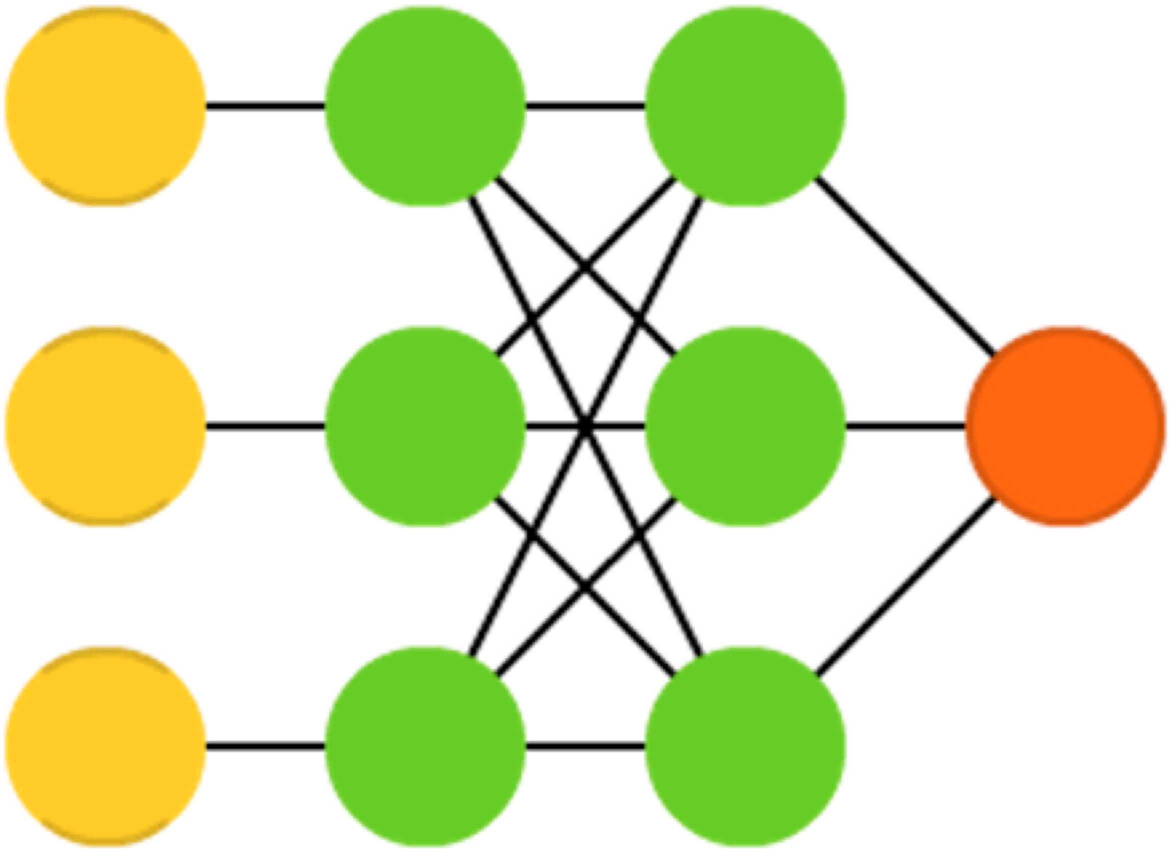
Generative Adversarial Networks (GAN)



Goodfellow, Ian, et al. "Generative adversarial nets." Advances in Neural Information Processing Systems. 2014.

Source: http://www.asimovinstitute.org/neural-network-zoo/

Support Vector Machines (SVM)

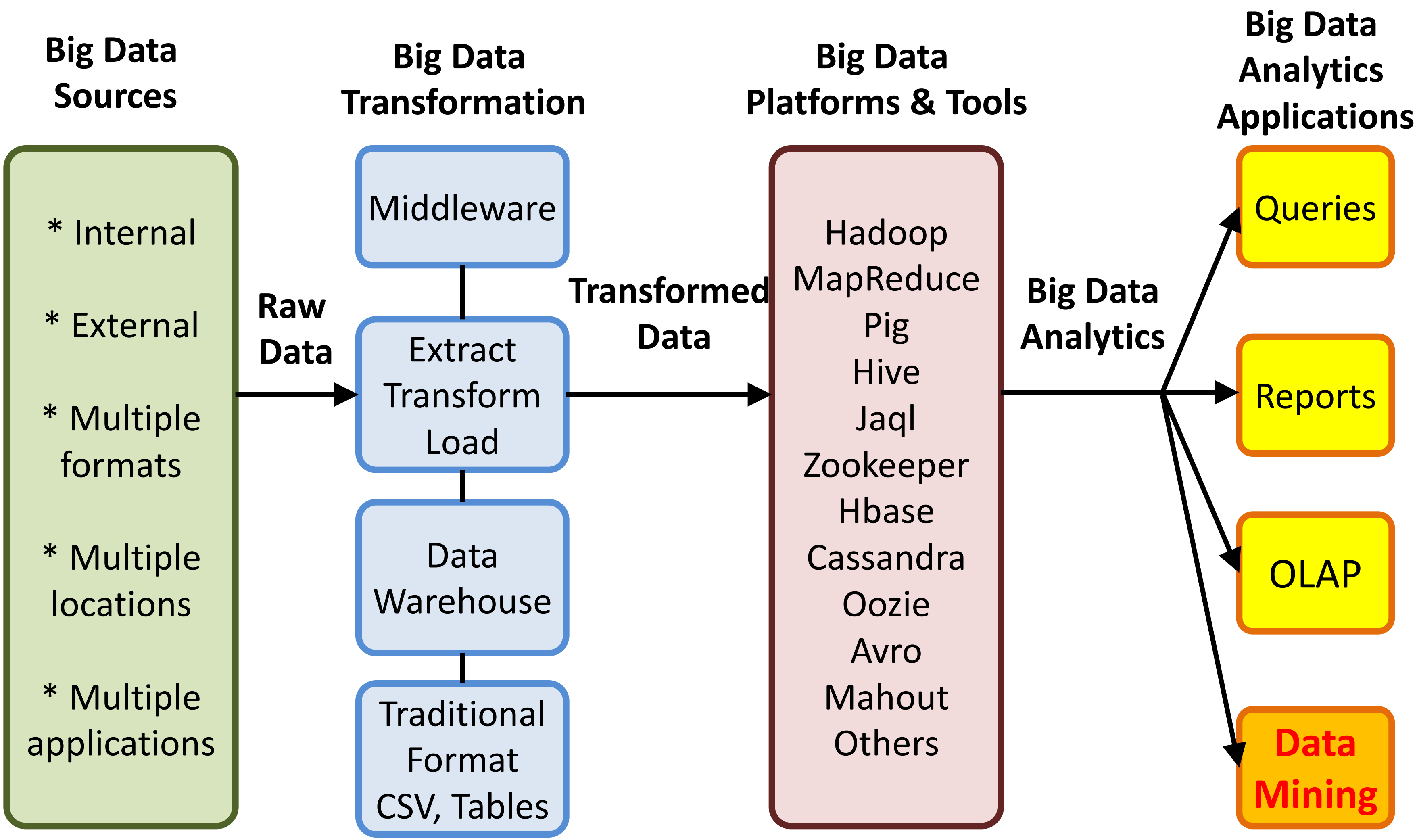


Cortes, Corinna, and Vladimir Vapnik. “Support-vector networks.” Machine learning 20.3 (1995): 273-297.

Source: http://www.asimovinstitute.org/neural-network-zoo/

Architectures of Big Data Analytics

Architecture of Big Data Analytics

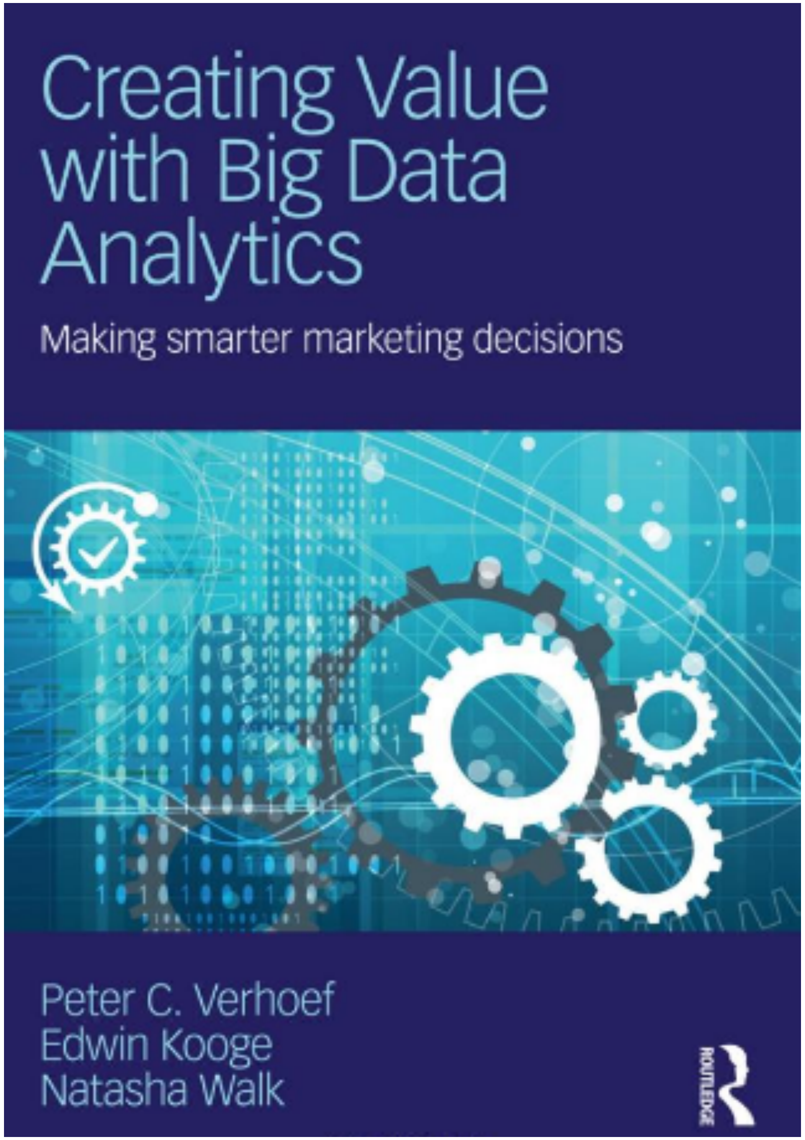


Source: Stephan Kudyba (2014), Big Data, Mining, and Analytics: Components of Strategic Decision Making, Auerbach Publications

Architecture of Big Data Analytics

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

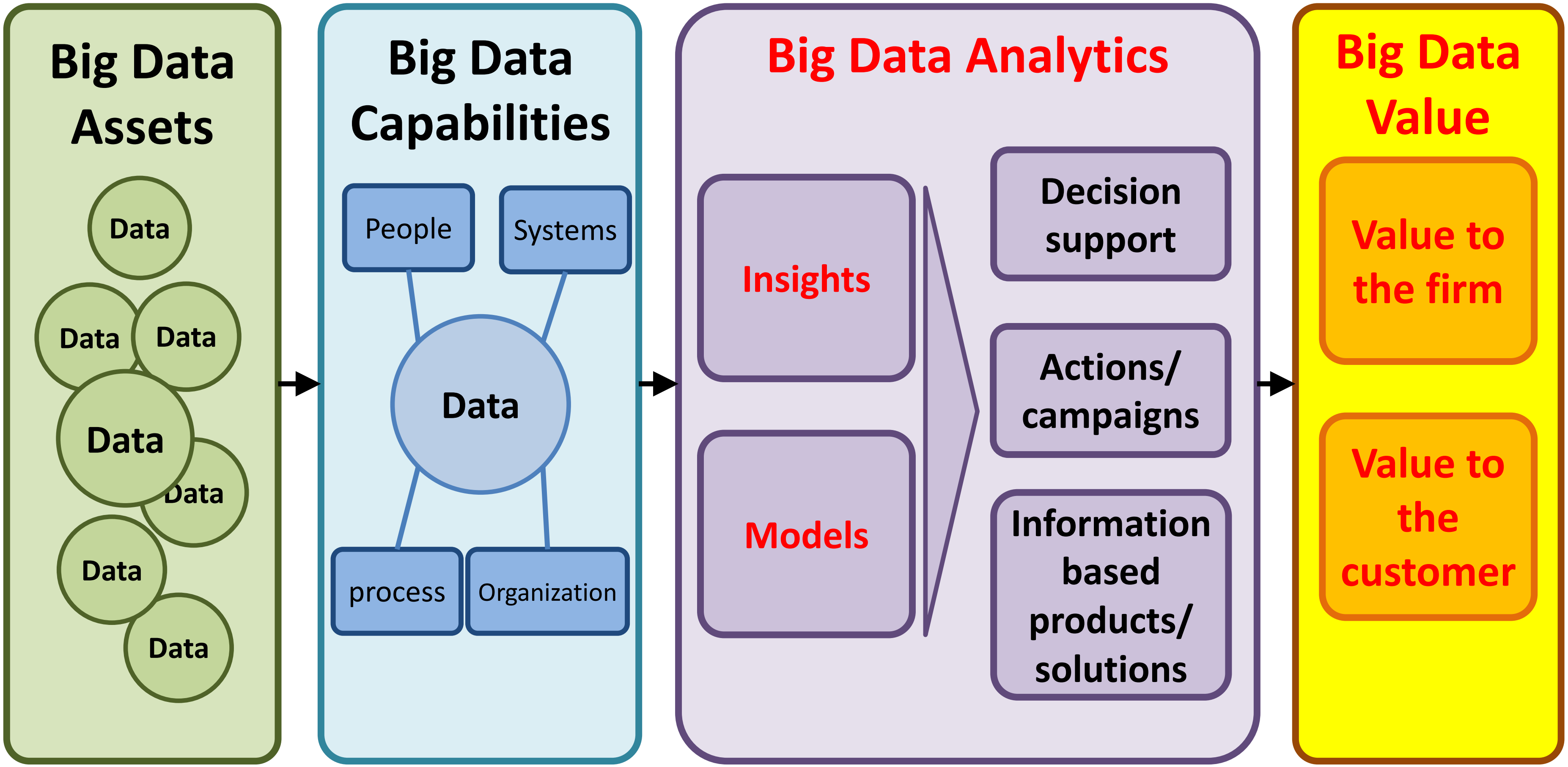
Creating Value with Big Data Analytics: Making Smarter Marketing Decisions, Peter C. Verhoef and Edwin Kooge, Routledge, 2016



Source: https://www.amazon.com/Creating-Value-Big-Data-Analytics/dp/1138837970

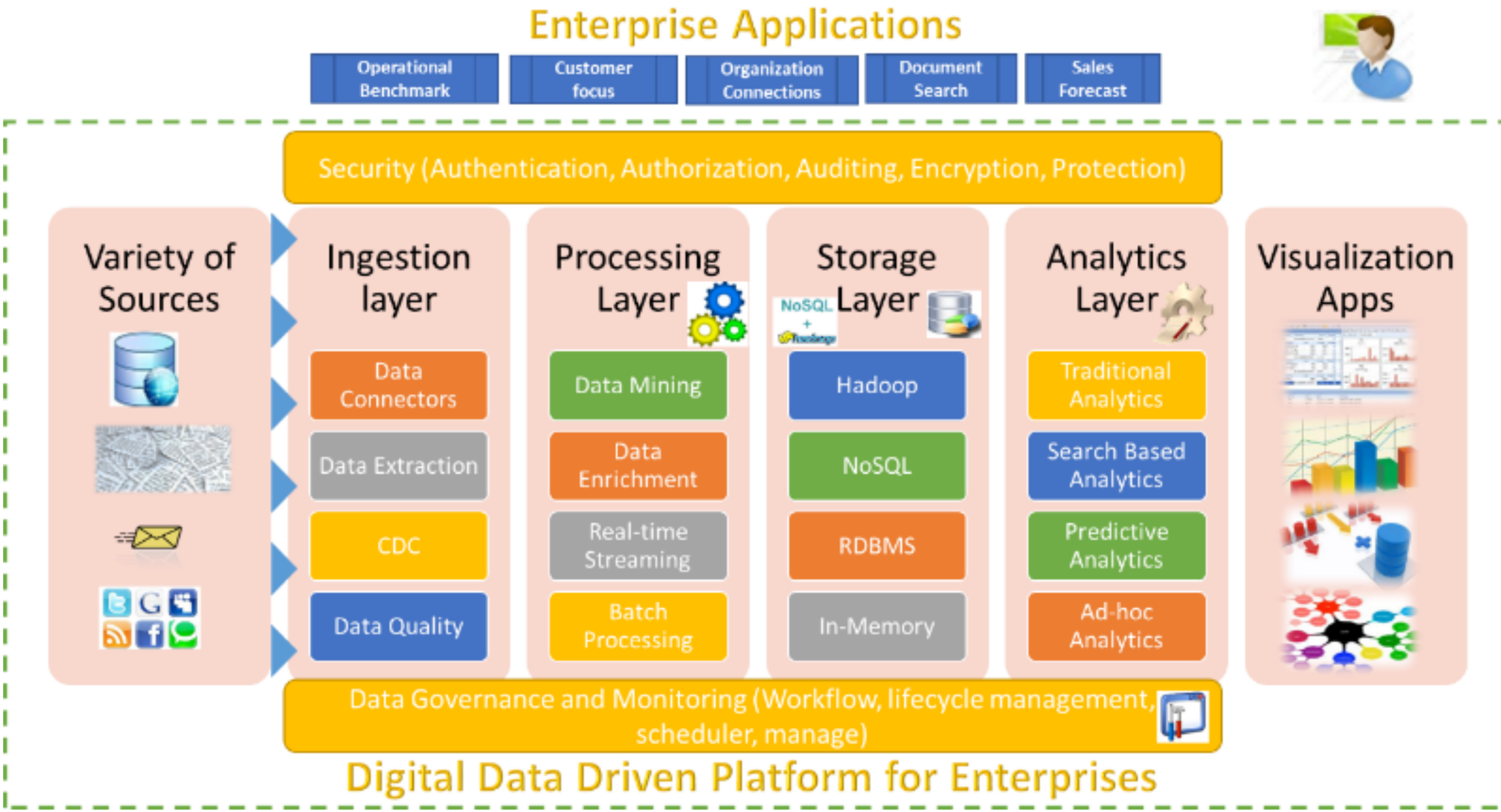
Big Data Value Creation Model

Creating Value with Big Data Analytics: Making Smarter Marketing Decisions



Source: Peter C. Verhoef and Edwin Kooge (2016), Creating Value with Big Data Analytics: Making Smarter Marketing Decisions, Routledge

Digital Data Platform for Enterprises
Big Data Analytics



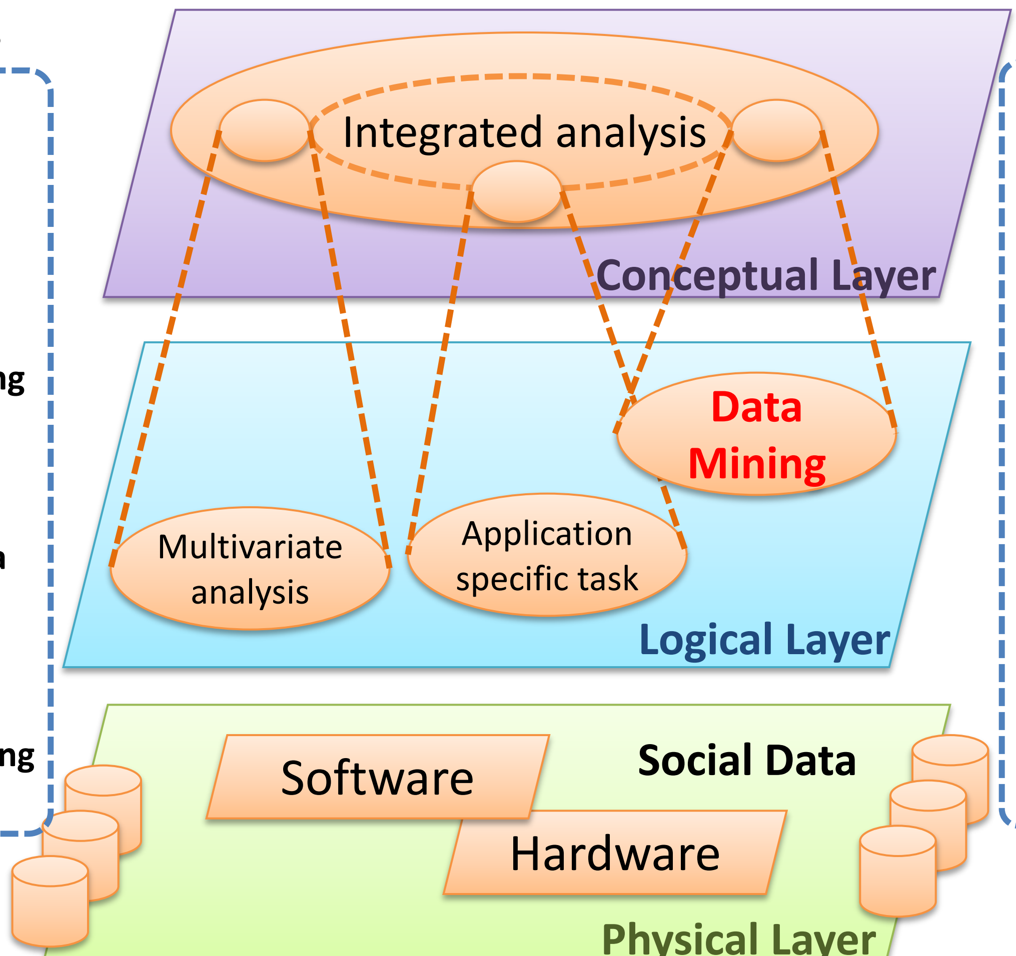
Source: https://blog.persistent.com/index.php/2015/05/05/is-your-enterprise-data-platform-ready-for-the-dive-into-digital-transformation/

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**

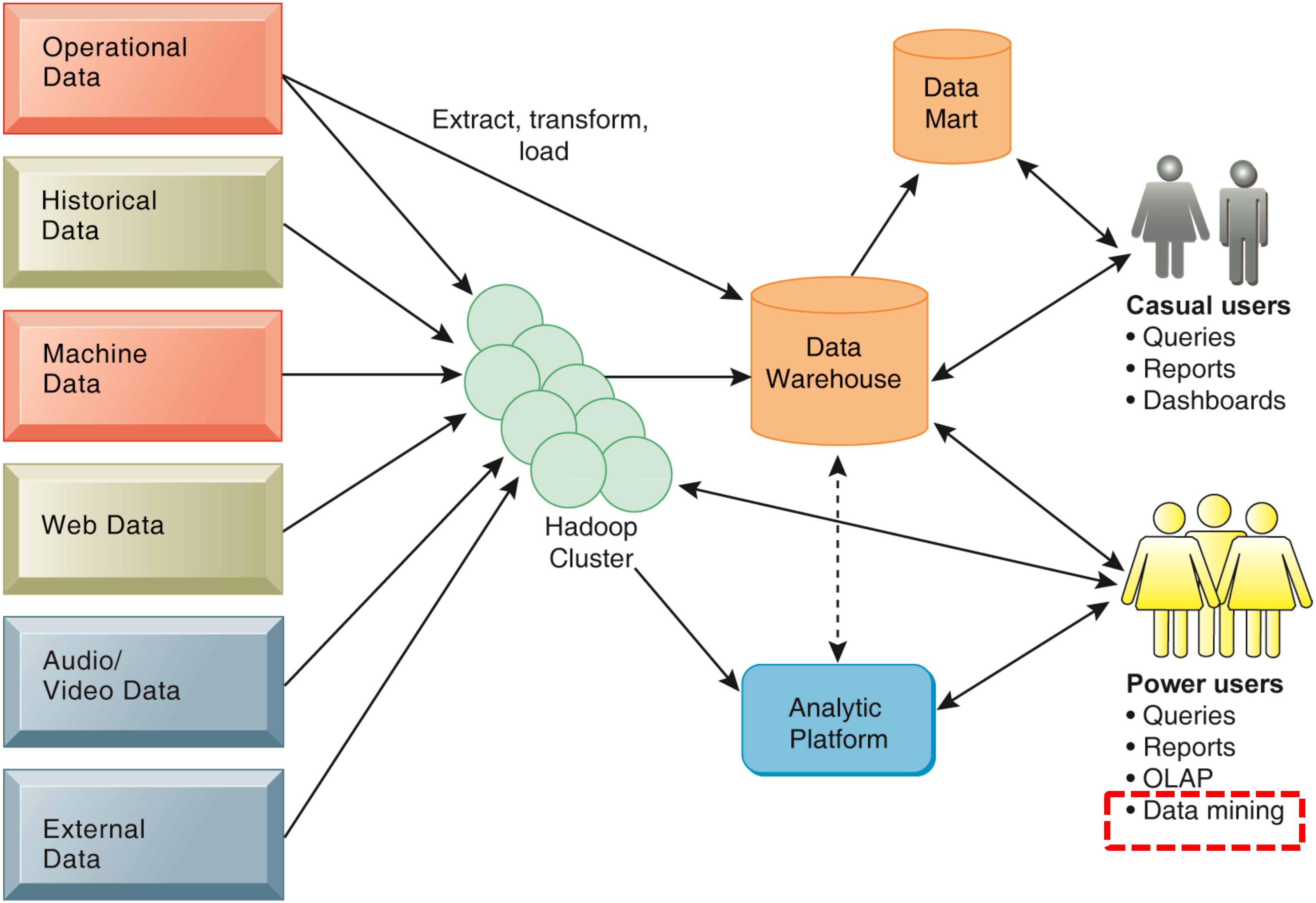


Analysts

- **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

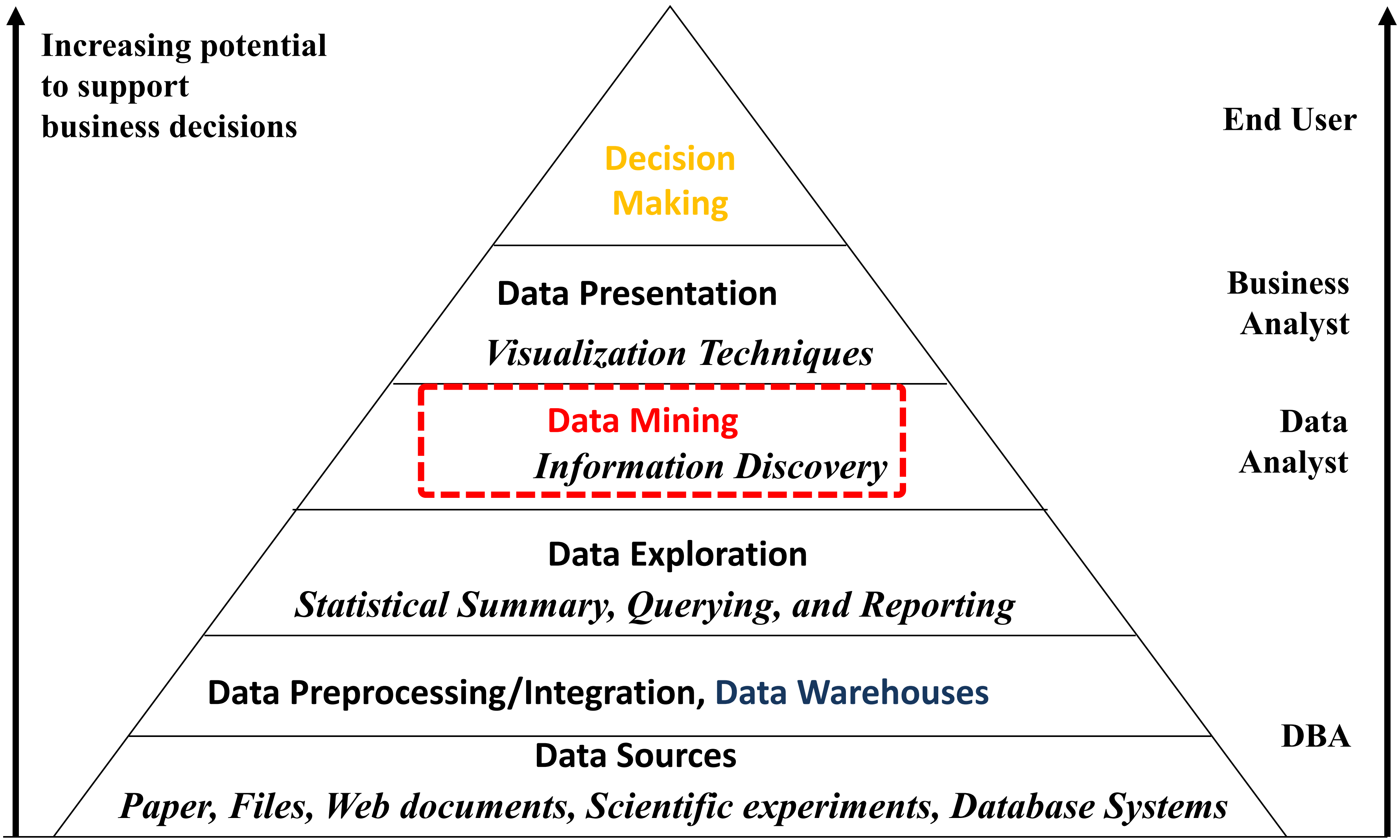
Business Intelligence (BI) Infrastructure



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

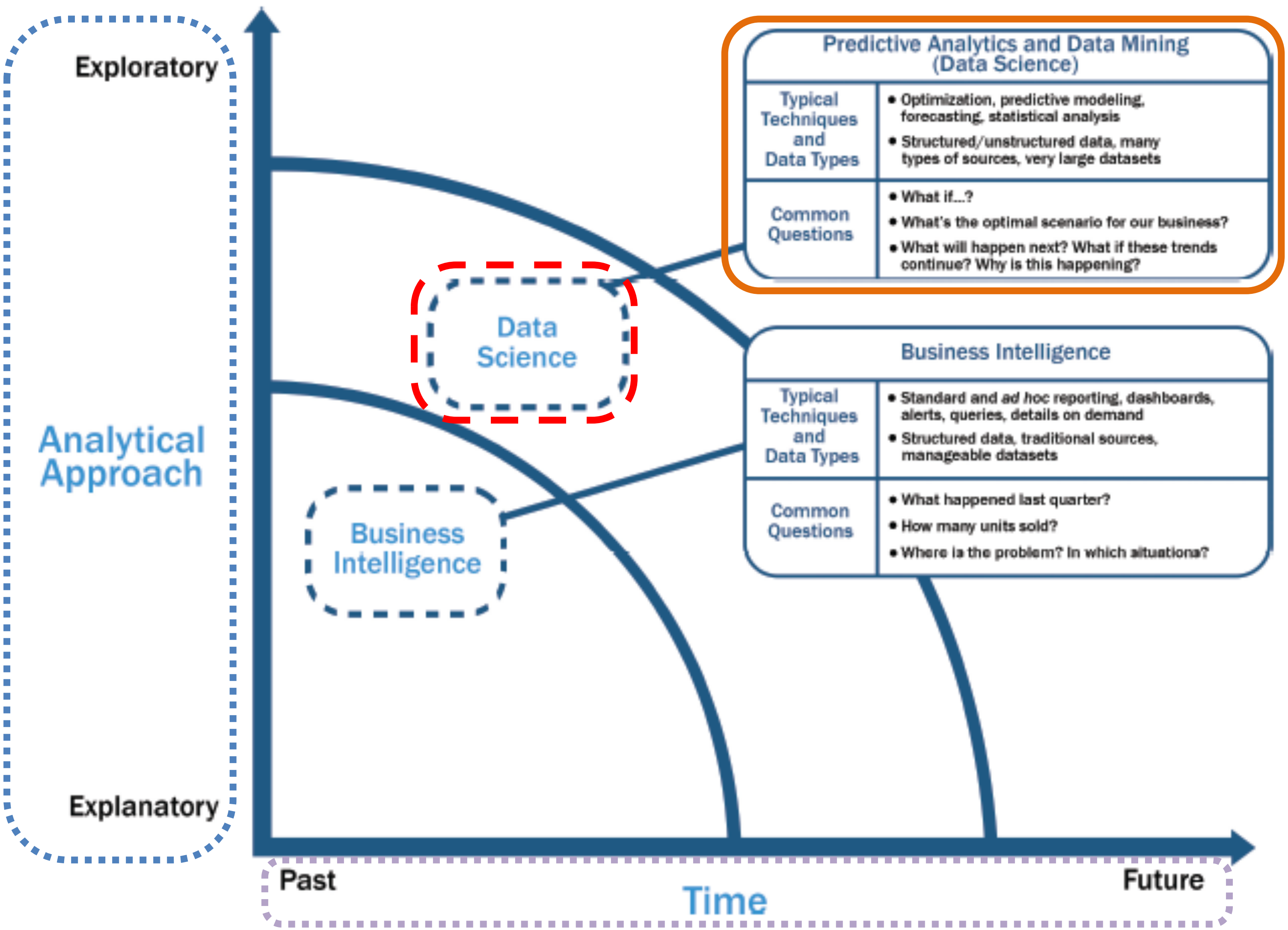


Source: Jiawei Han and Micheline Kamber (2006), Data Mining: Concepts and Techniques, Second Edition, Elsevier

The Evolution of BI Capabilities

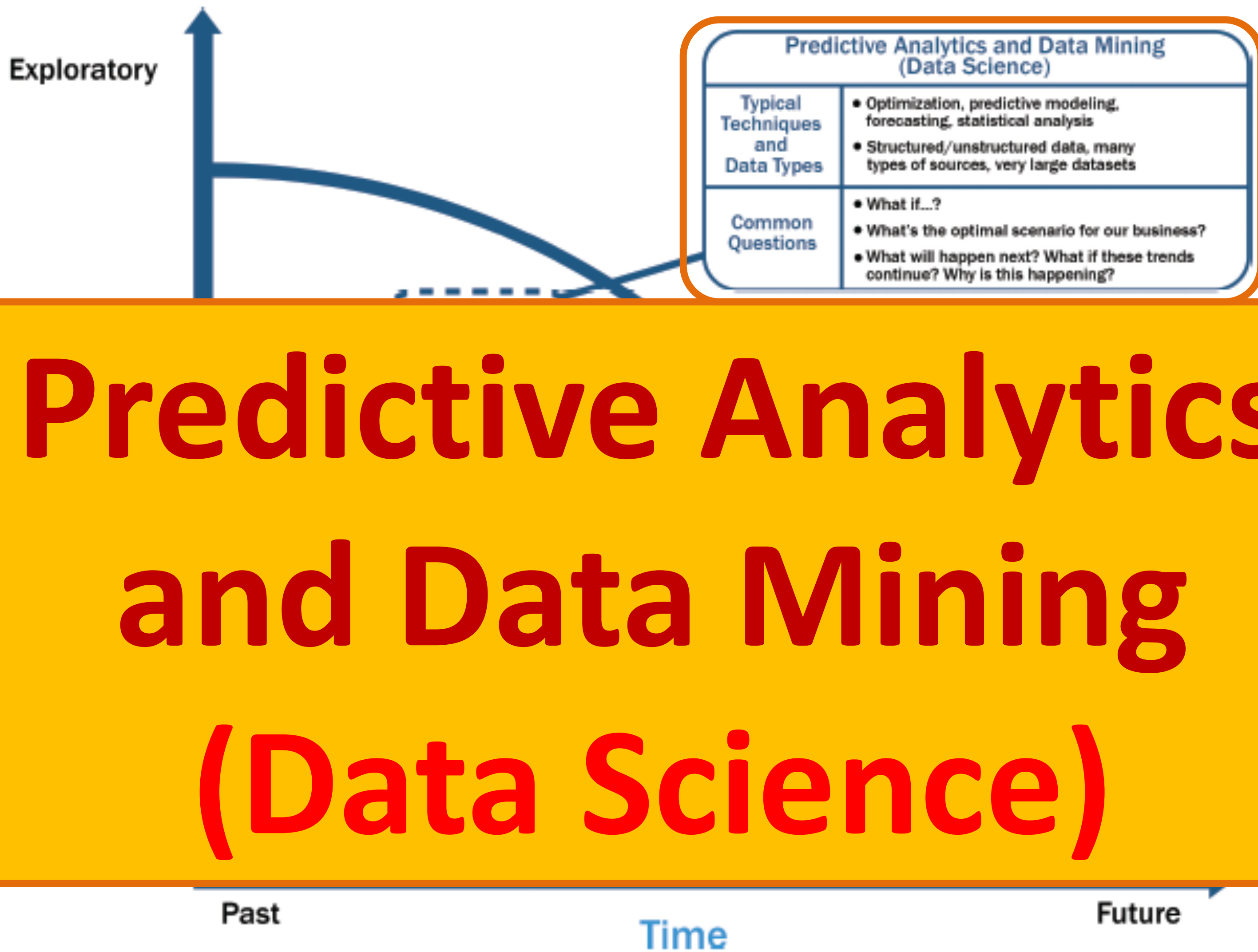
Source: Turban et al. (2011), Decision Support and Business Intelligence Systems

Data Science and Business Intelligence



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)

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?

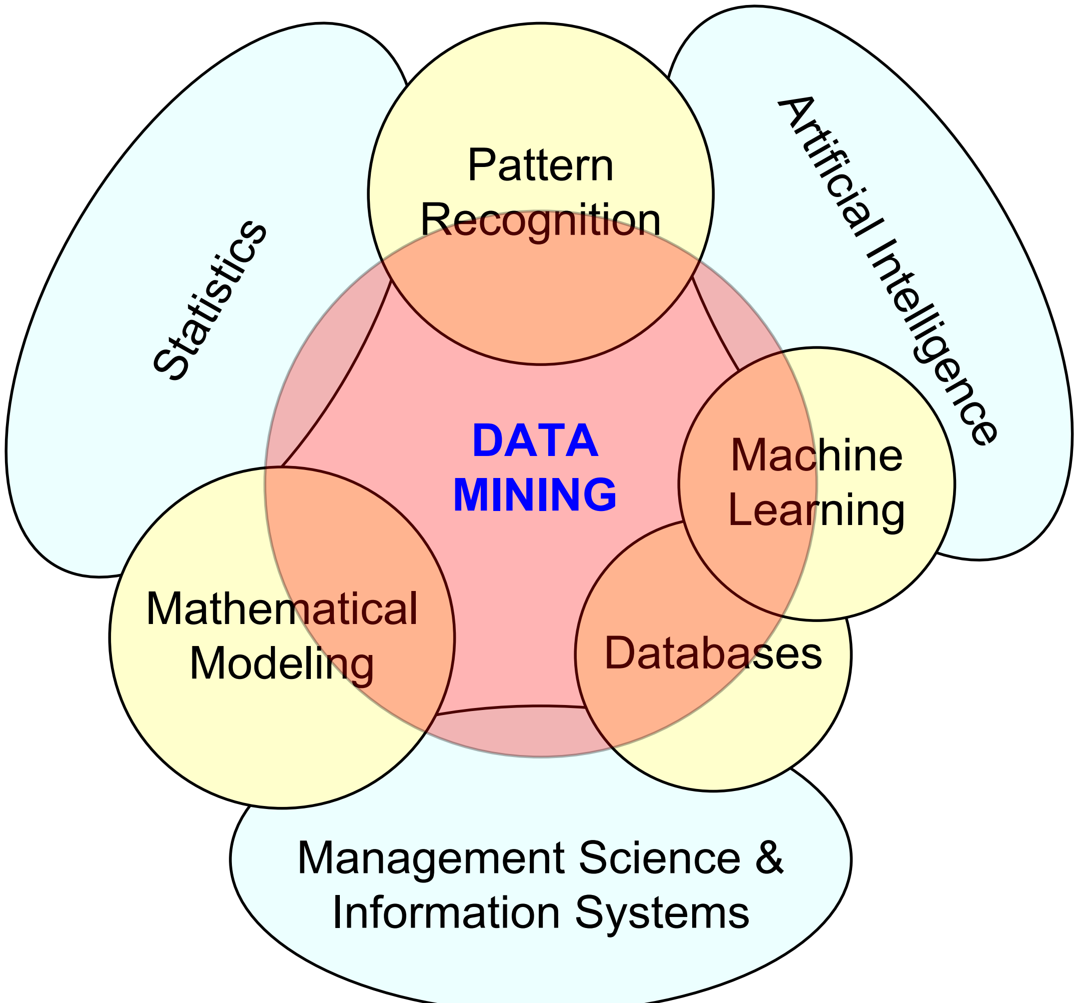
Source: EMC Education Services, Data Science and Big Data Analytics: Discovering, Analyzing, Visualizing and Presenting Data, Wiley, 2015

Data Mining

AI

Machine Learning

Data Mining at the Intersection of Many Disciplines



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

Data Mining

Advanced Data Analysis

Evolution of
Database System Technology

Evolution of Database System Technology

Data Collection and Database Creation

(1960s and earlier)

- Primitive file processing

Database Management Systems

(1970s–early 1980s)

- Hierarchical and network database systems
- Relational database systems
- Query languages: SQL, etc.
- Transactions, concurrency control and recovery
- On-line transaction processing (OLTP)

Advanced Database Systems

(mid-1980s–present)

- Advanced data models: extended relational, object-relational, etc.
- Advanced applications: spatial, temporal, multimedia, active, stream and sensor, scientific and engineering, knowledge-based
- XML-based database systems
- Integration with information retrieval
- Data and information integration

Advanced Data Analysis:

(late 1980s–present)

- Data warehouse and OLAP
- Data mining and knowledge discovery: generalization, classification, association, clustering
- Advanced data mining applications: stream data mining, bio-data mining, time-series analysis, text mining, Web mining, intrusion detection, etc.
- Data mining applications
- Data mining and society

New Generation of Information Systems

(present–future)

Source: Jiawei Han and Micheline Kamber (2011), Data Mining: Concepts and Techniques, Third Edition, Elsevier

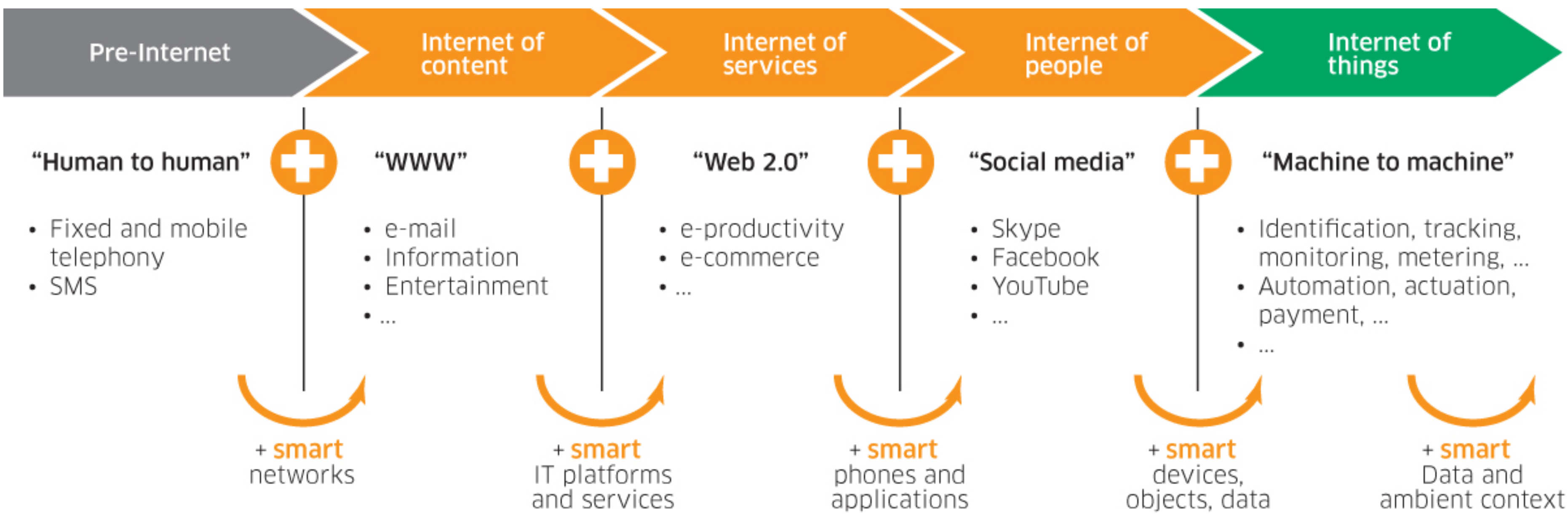
Big Data Analysis

- **Too Big,**
  too Unstructured,
  too many different source
  to be manageable through
  traditional databases

Internet Evolution

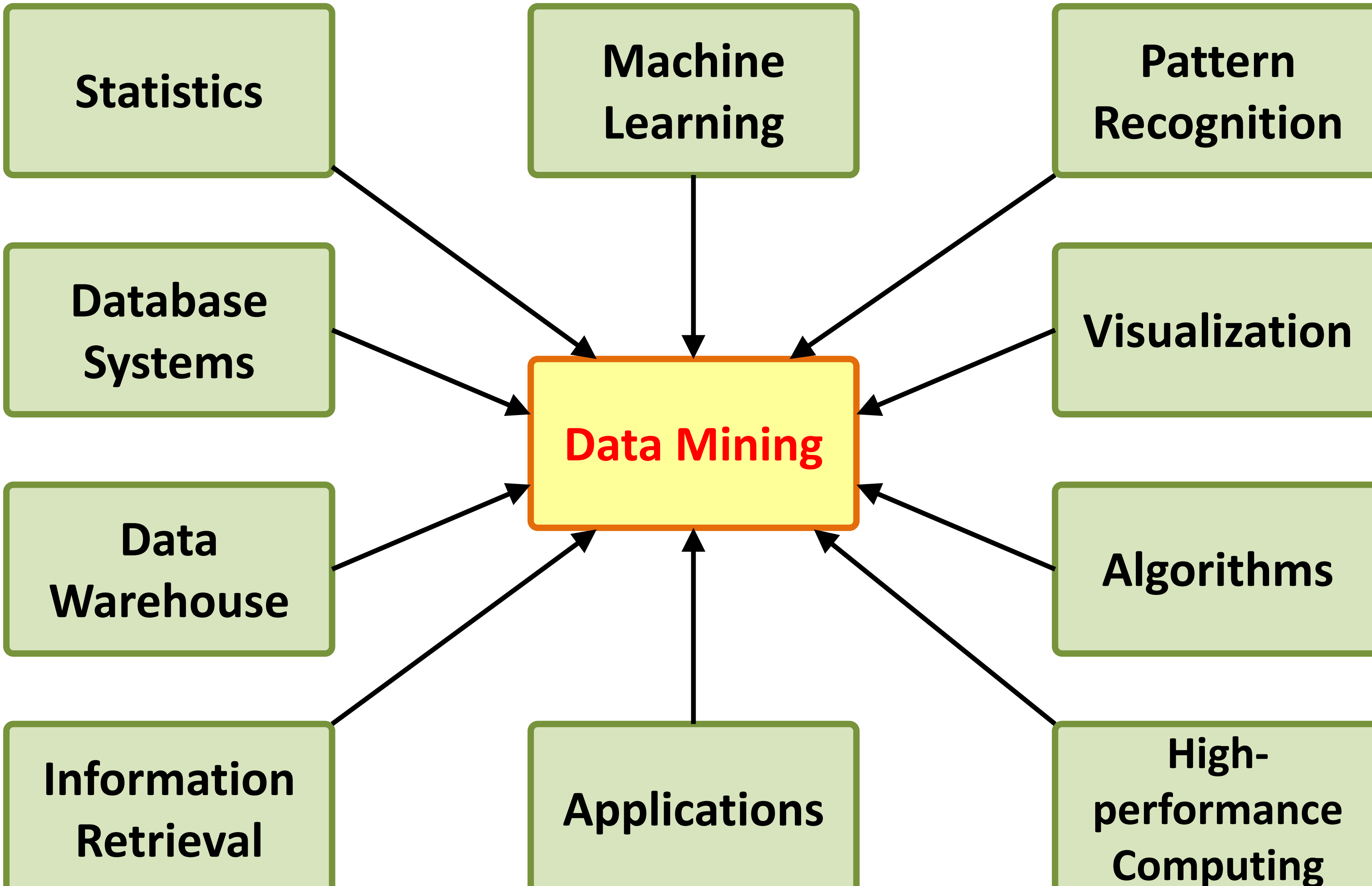
Internet of People (IoP): Social Media

Internet of Things (IoT): Machine to Machine



Source: Marc Jadoul (2015), The IoT: The next step in internet evolution, March 11, 2015
http://www2.alcatel-lucent.com/techzine/iot-internet-of-things-next-step-evolution/

Data Mining Technologies



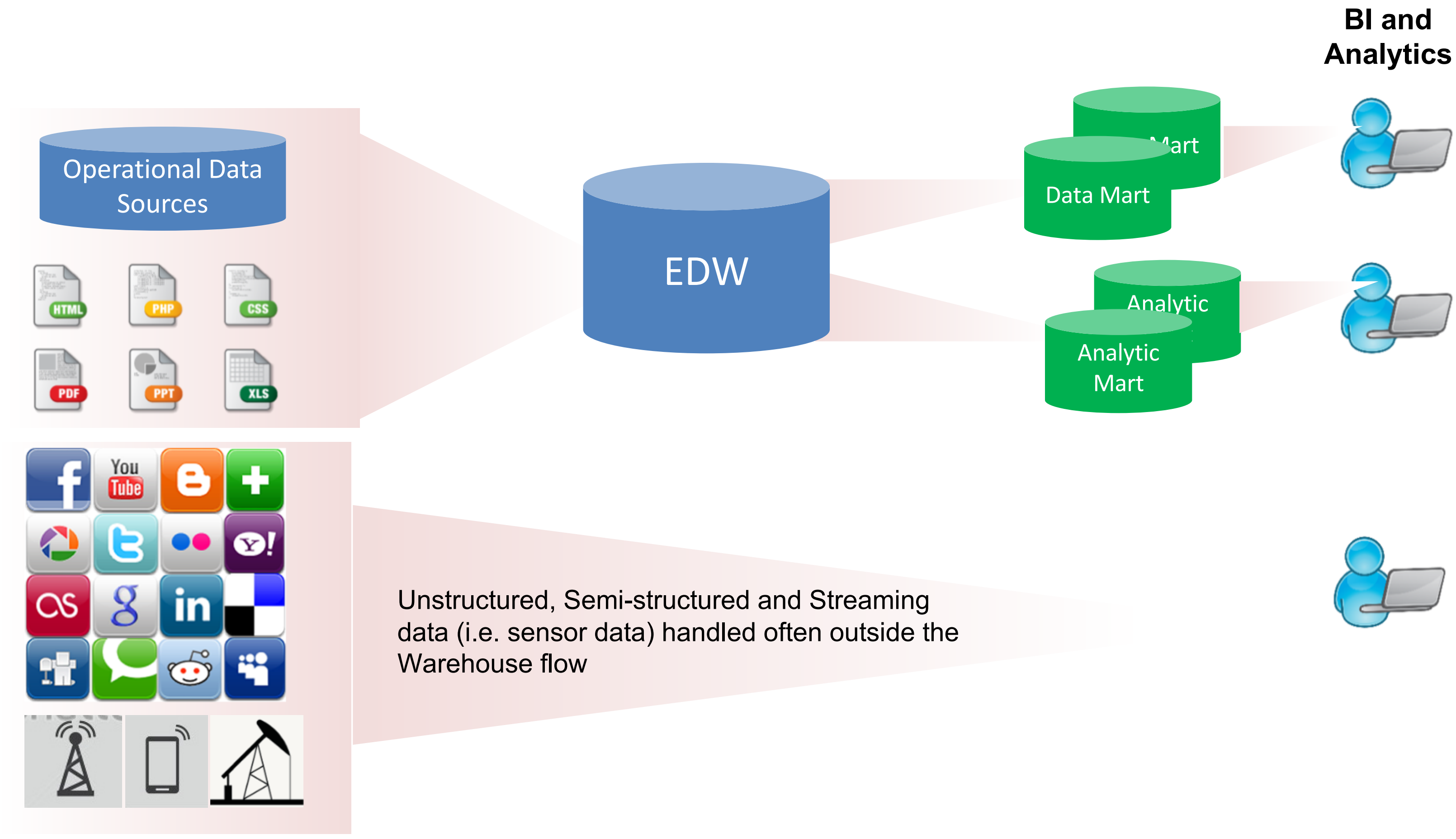
Source: Jiawei Han and Micheline Kamber (2011), Data Mining: Concepts and Techniques, Third Edition, Elsevier

A Taxonomy for Data Mining Tasks

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

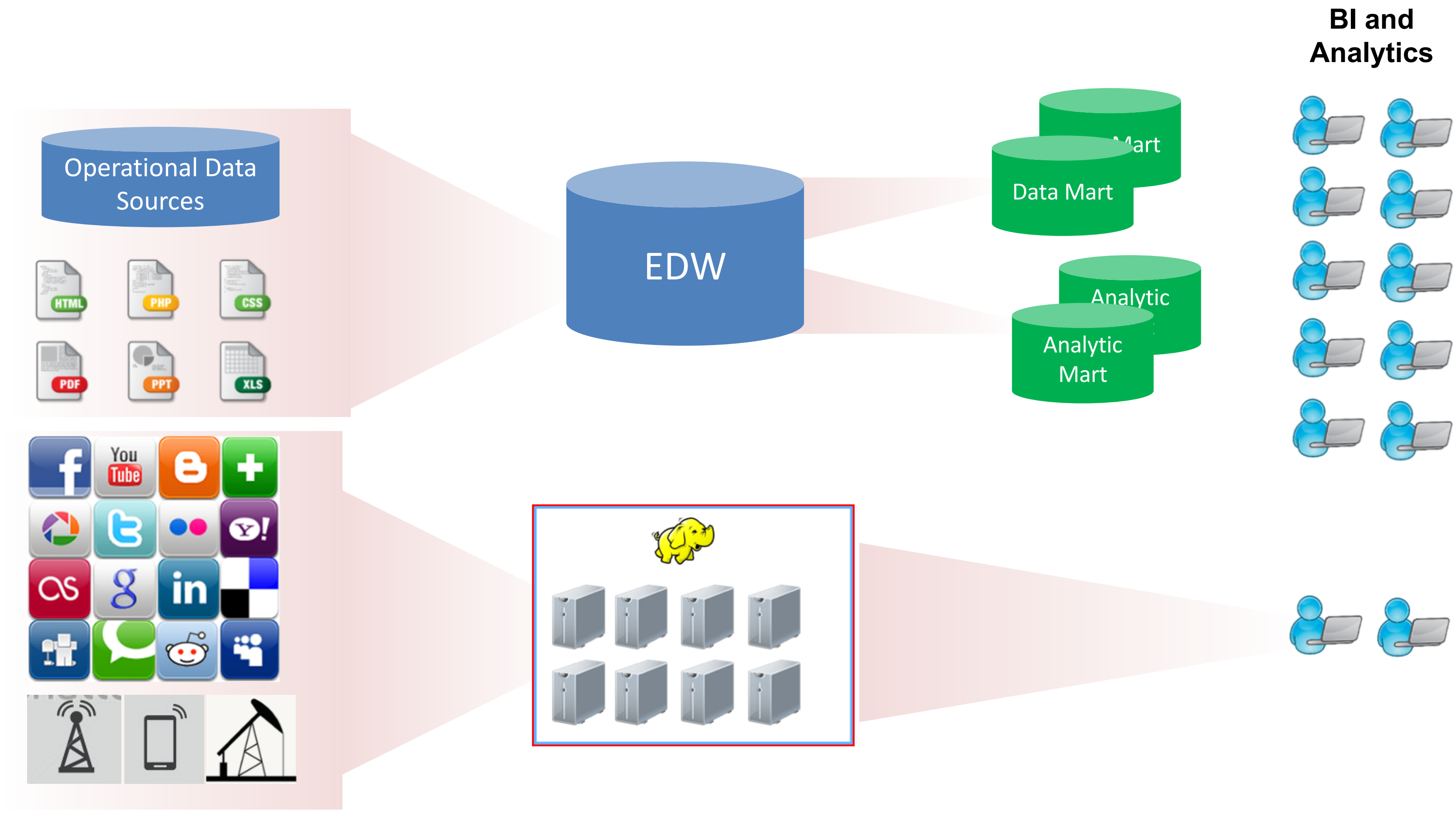
Source: Turban et al. (2011), Decision Support and Business Intelligence Systems

Traditional Analytics



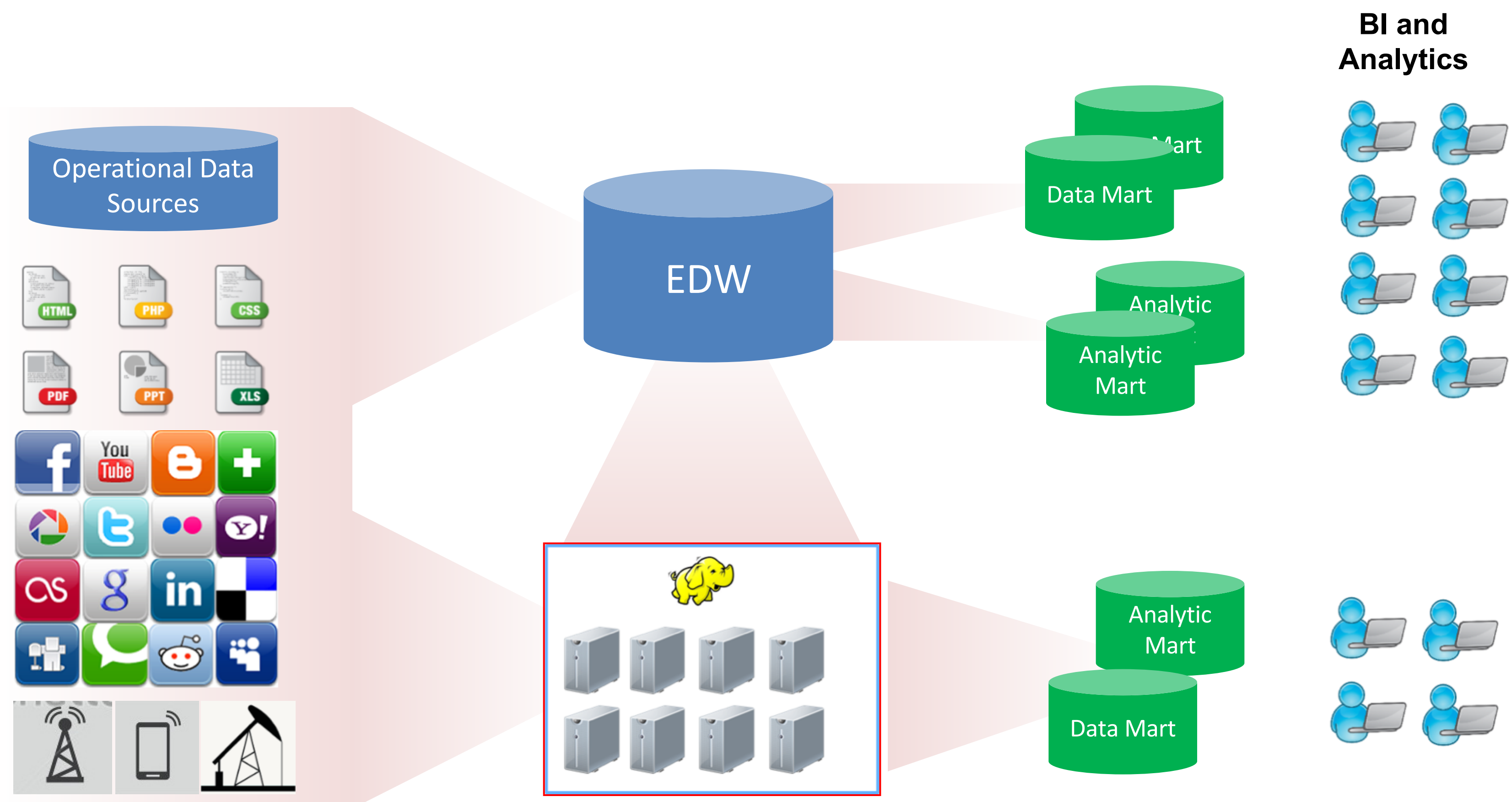
Source: Deepak Ramanathan (2014), SAS Modernization architectures - Big Data Analytics

Hadoop as a “new data” Store



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

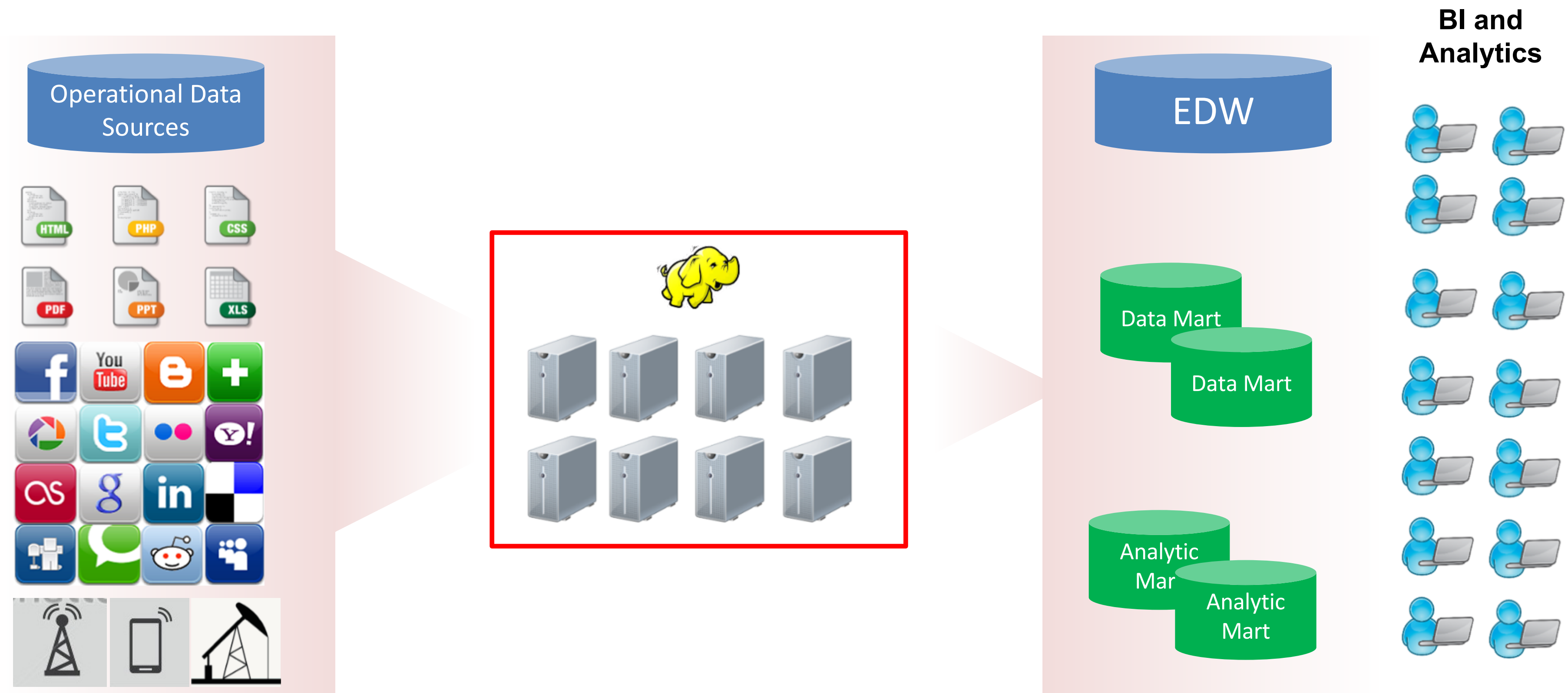
Hadoop as an additional input to the EDW



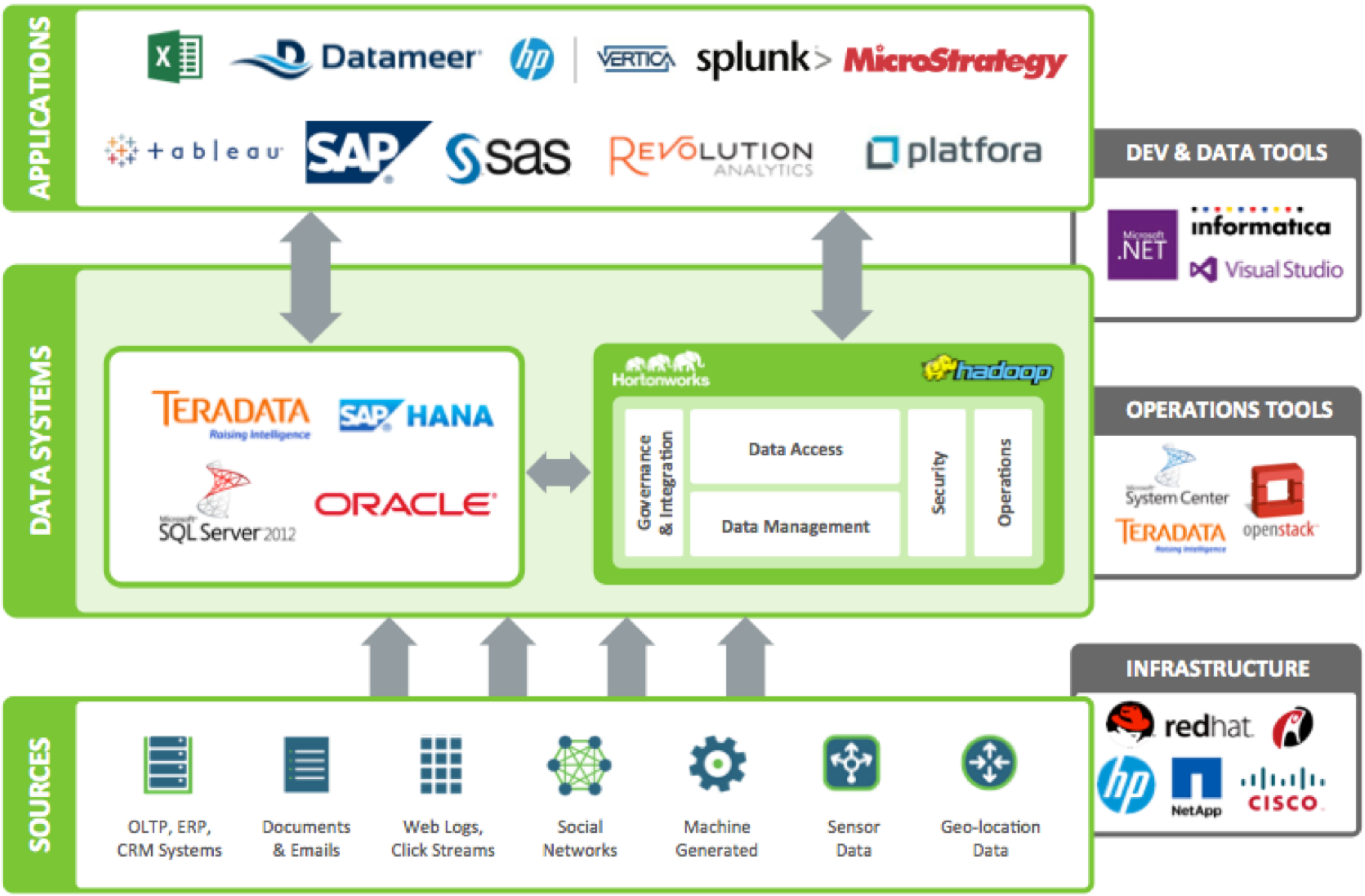
Source: Deepak Ramanathan (2014), SAS Modernization architectures - Big Data Analytics

Hadoop Data Platform As a “staging Layer” as part of a “data Lake”

– Downstream stores could be Hadoop, data appliances or an RDBMS

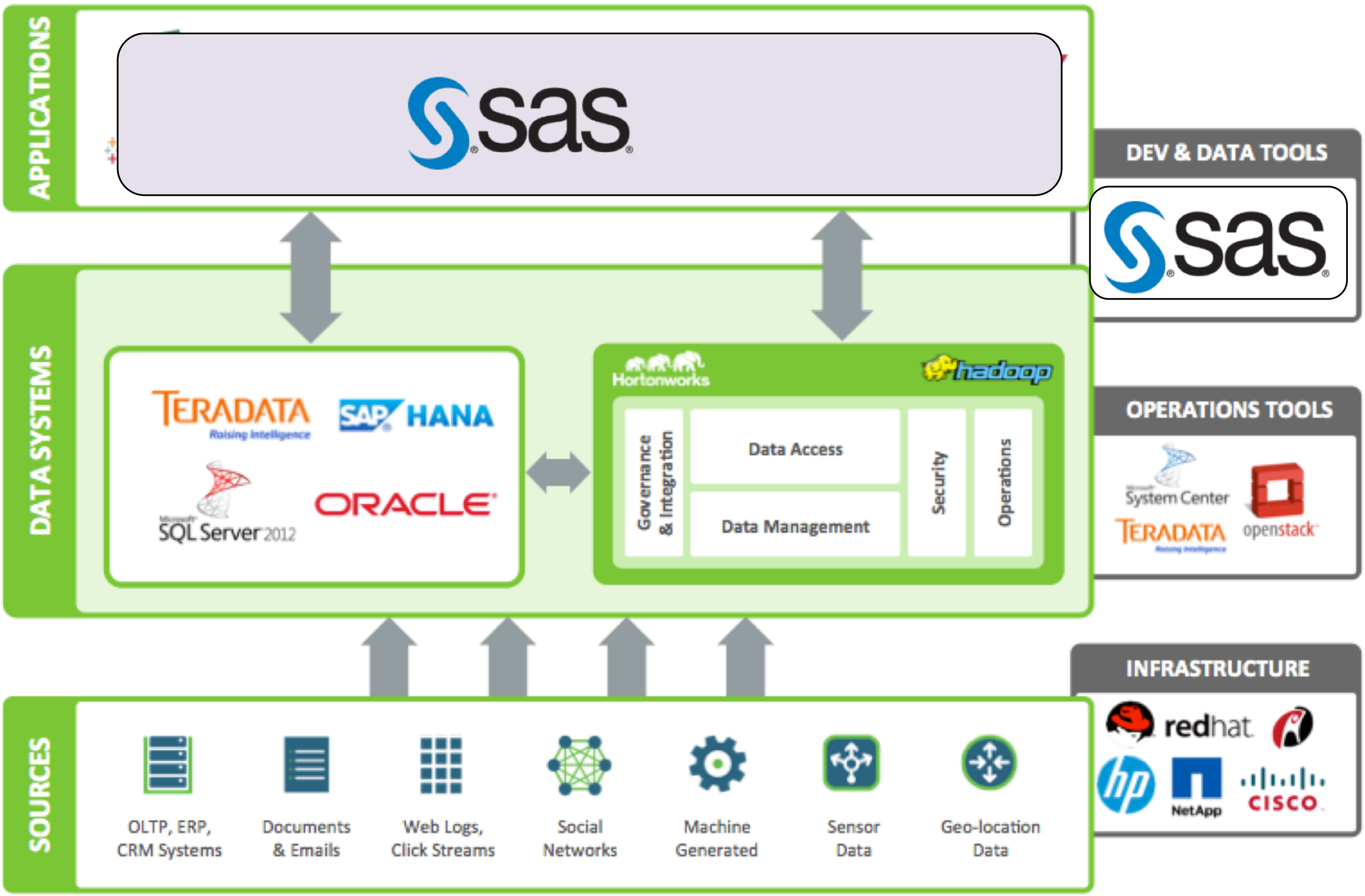


SAS Big data Strategy – SAS areas



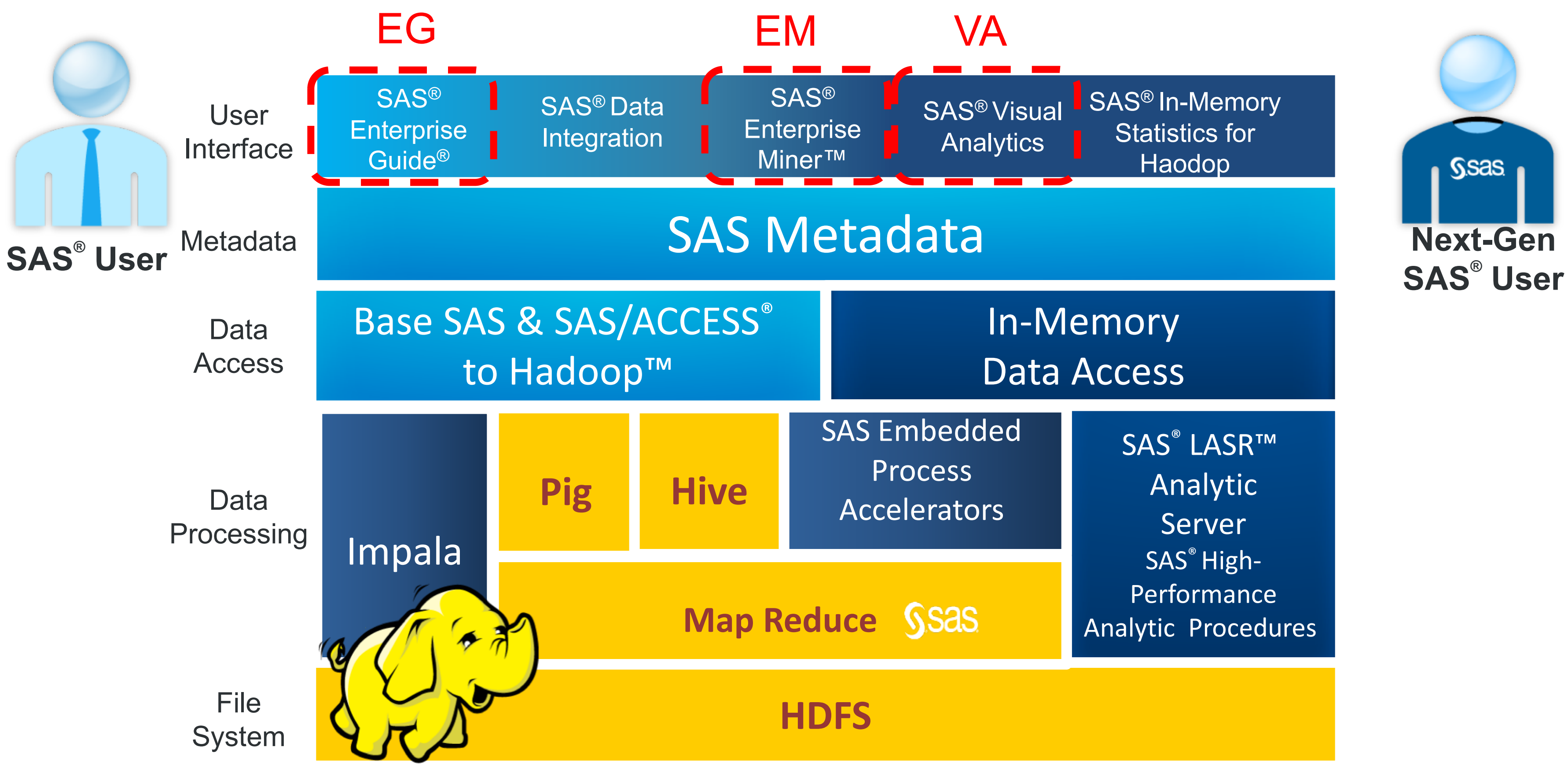
Source: Deepak Ramanathan (2014), SAS Modernization architectures - Big Data Analytics

SAS Big data Strategy – SAS areas



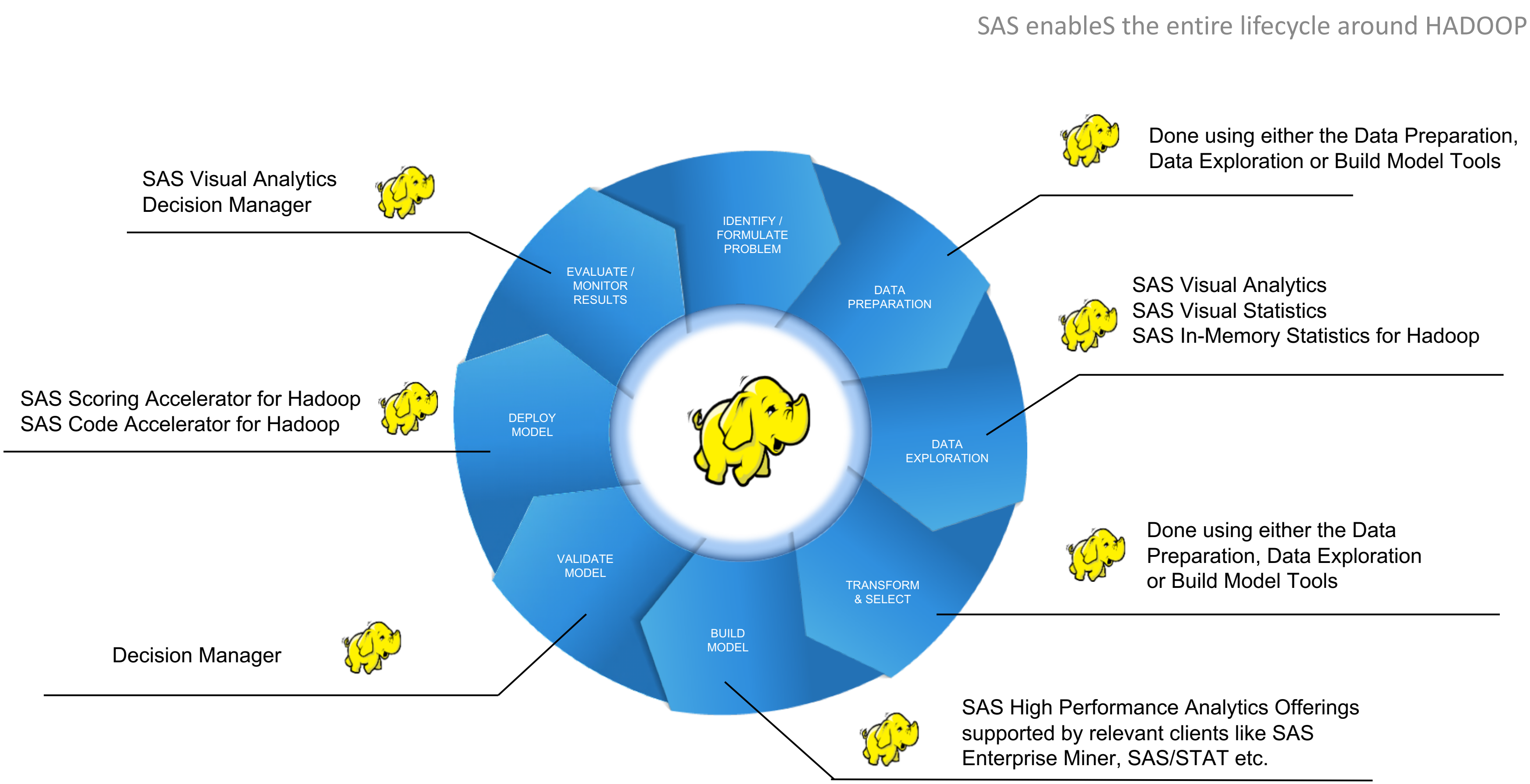
Source: Deepak Ramanathan (2014), SAS Modernization architectures - Big Data Analytics

SAS® Within the HADOOP ECOSYSTEM



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

SAS enables the entire lifecycle around HADOOP



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

Big Data,
Big Analytics:

Emerging Business Intelligence and Analytic Trends for Today's Businesses

Big Data,
Prediction
vs.
Explanation

Source: Agarwal, R., & Dhar, V. (2014). Editorial—Big Data, Data Science, and Analytics: The Opportunity and Challenge for IS Research. Information Systems Research, 25(3), 443-448.

Big Data:

The Management Revolution

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

Business Intelligence and Enterprise Analytics

- Predictive analytics
- Data mining
- Business analytics
- Web analytics
- Big-data analytics

Source: Thomas H. Davenport, "Enterprise Analytics: Optimize Performance, Process, and Decisions Through Big Data", FT Press, 2012

Three Types of Business Analytics

- Prescriptive Analytics
- Predictive Analytics
- Descriptive Analytics

Source: Thomas H. Davenport, "Enterprise Analytics: Optimize Performance, Process, and Decisions Through Big Data", FT Press, 2012

Three Types of Business Analytics

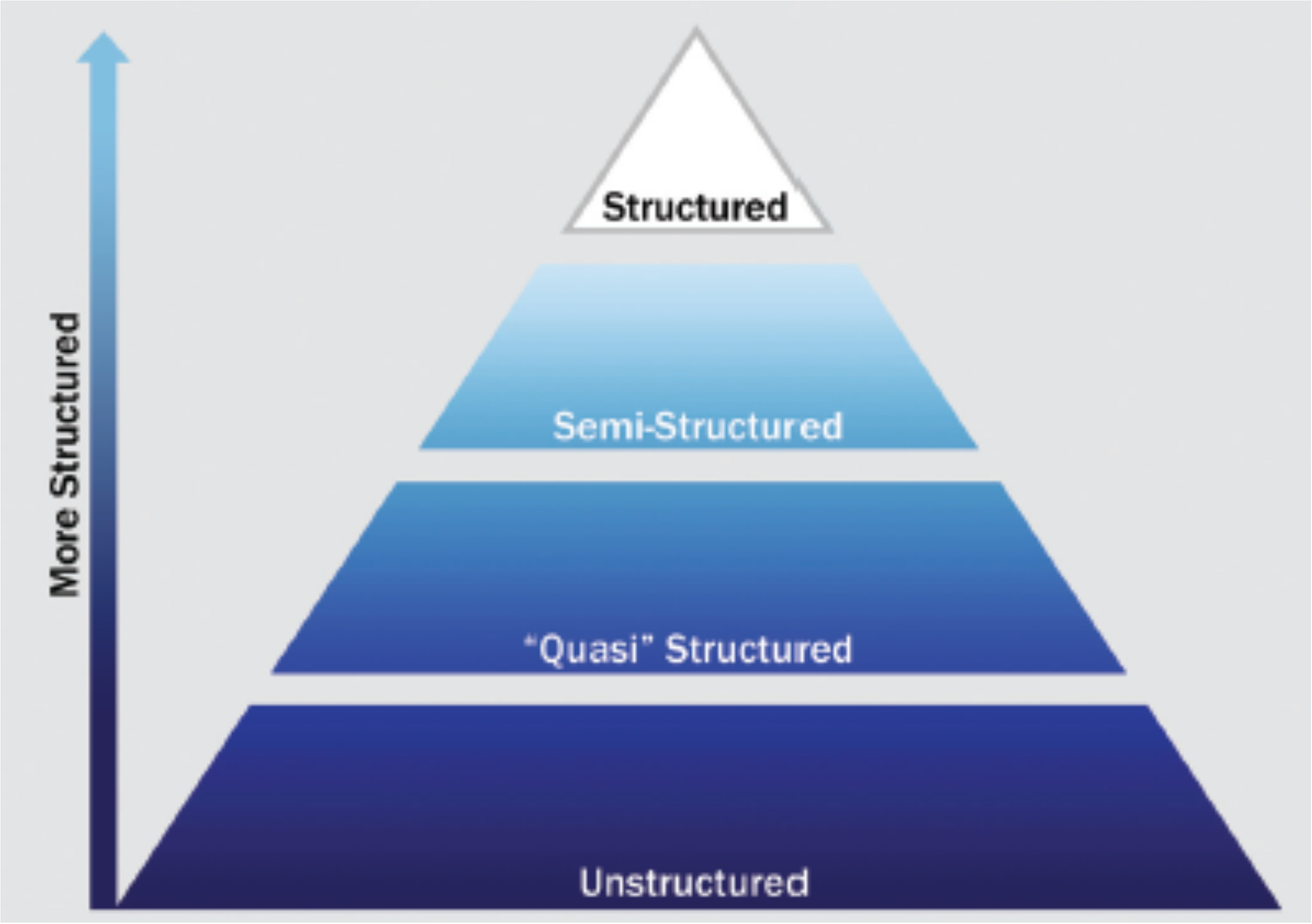
Optimization	"What's the best that can happen?"	Prescriptive Analytics
Randomized Testing	"What if we try this?"	Prescriptive Analytics
Predictive Modeling / Forecasting	"What will happen next?"	Predictive Analytics
Statistical Modeling	"Why is this happening?"	Predictive Analytics
Alerts	"What actions are needed?"	Descriptive Analytics
Query / Drill Down	"What exactly is the problem?"	Descriptive Analytics
Ad hoc Reports / Scorecards	"How many, how often, where?"	Descriptive Analytics
Standard Report	"What happened?"	Descriptive Analytics

Source: Thomas H. Davenport, "Enterprise Analytics: Optimize Performance, Process, and Decisions Through Big Data", FT Press, 2012

Big Data

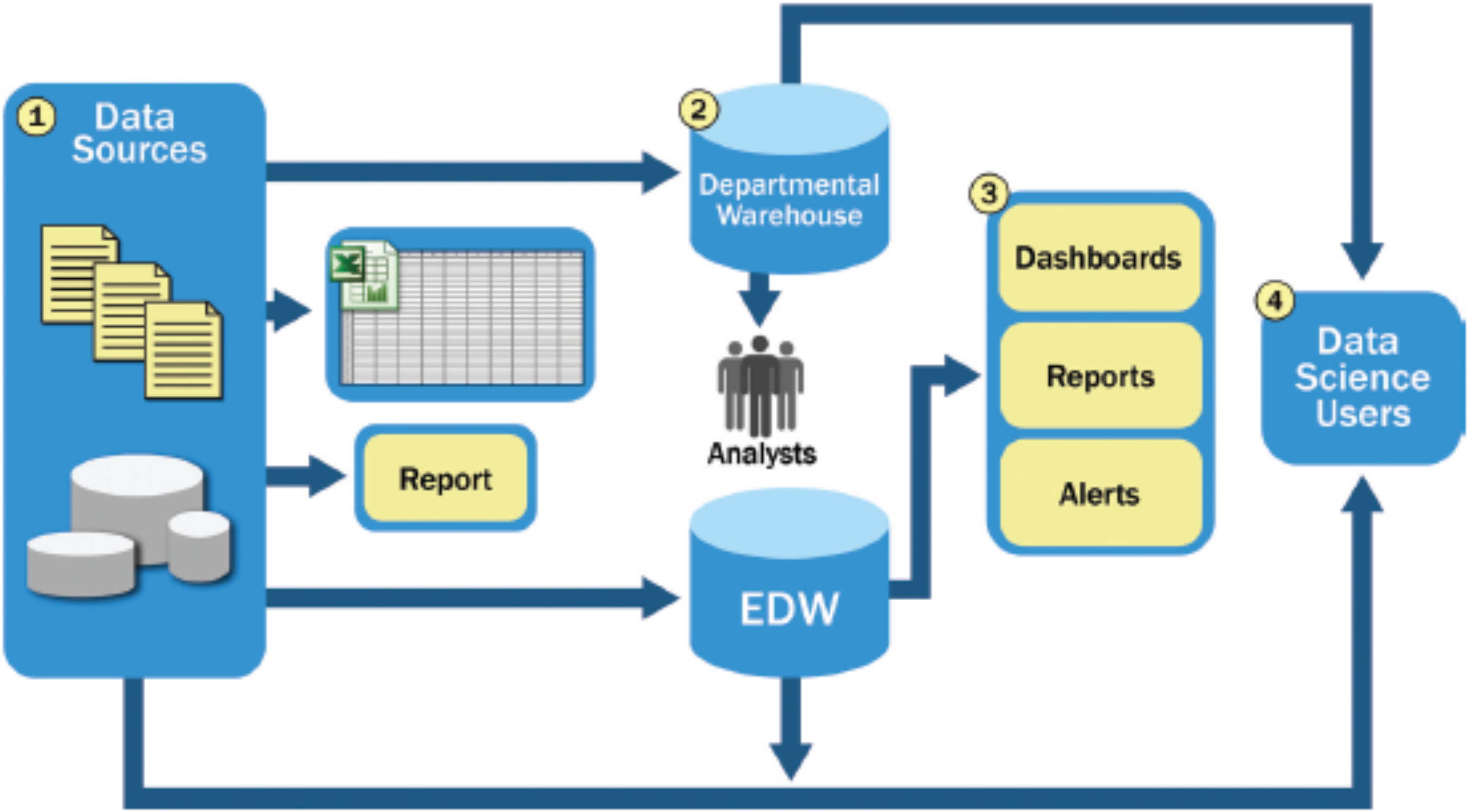
Source: EMC Education Services, Data Science and Big Data Analytics: Discovering, Analyzing, Visualizing and Presenting Data, Wiley, 2015

Big Data Growth is increasingly unstructured



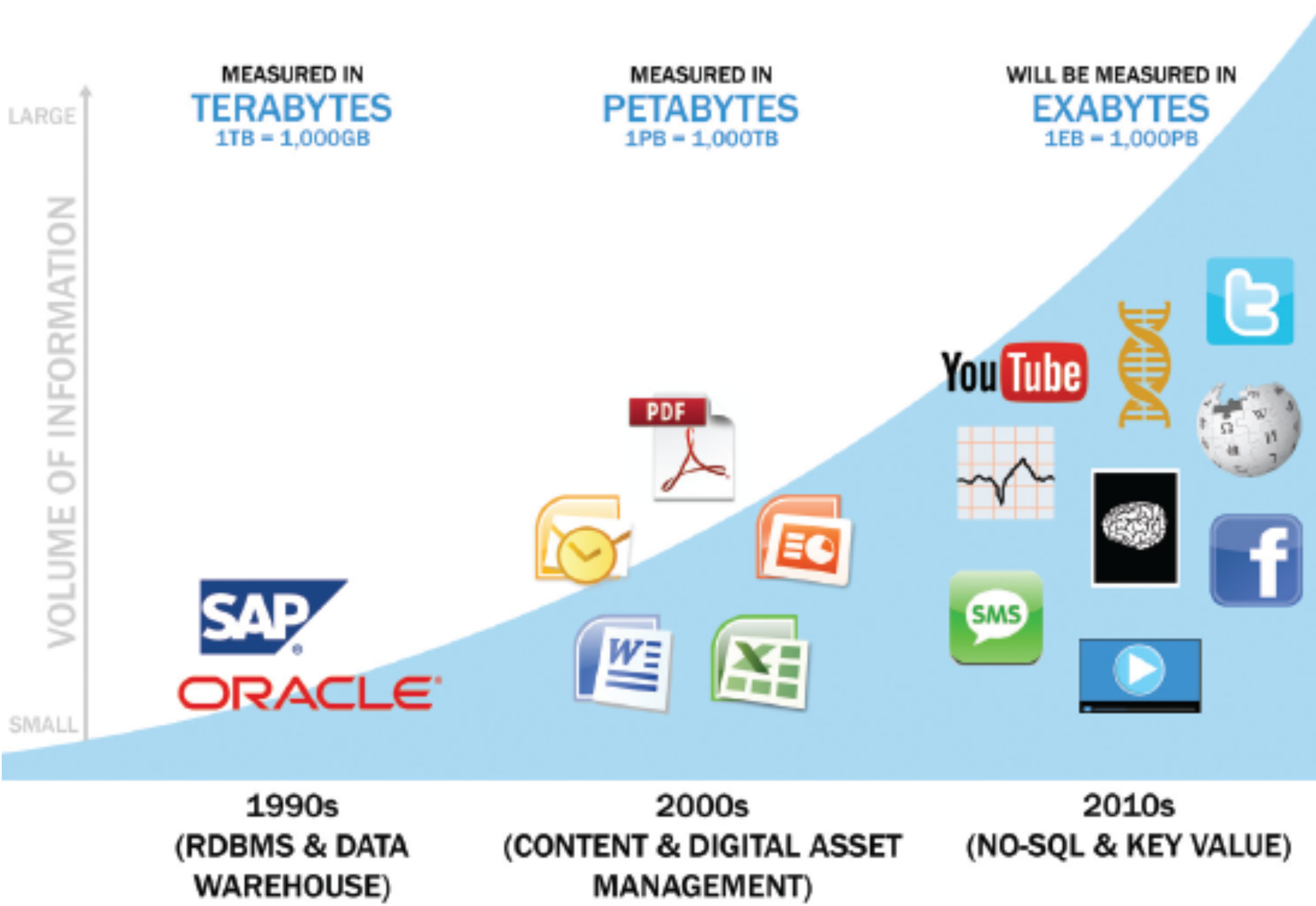
Source: EMC Education Services, Data Science and Big Data Analytics: Discovering, Analyzing, Visualizing and Presenting Data, Wiley, 2015

Typical Analytic Architecture



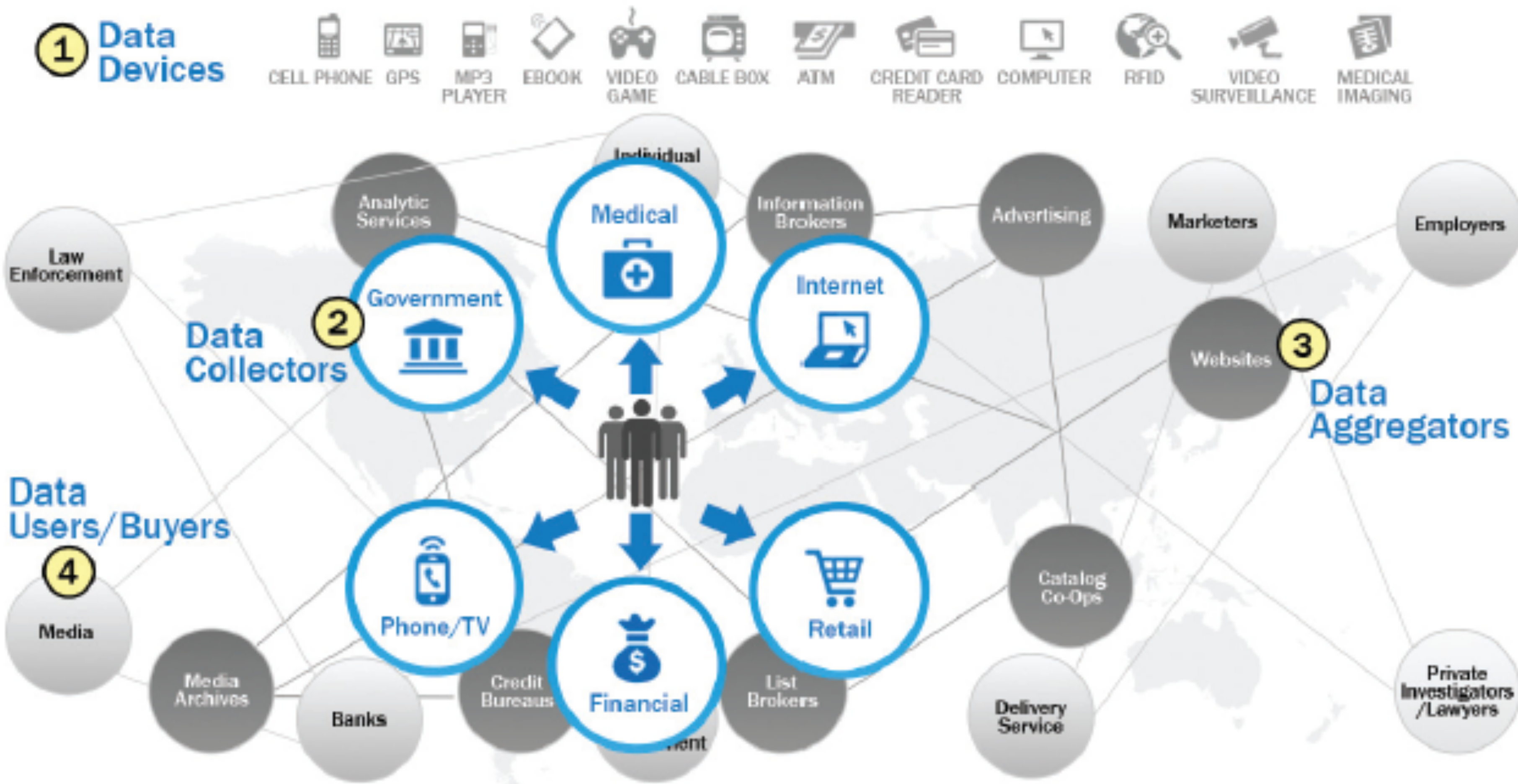
Source: EMC Education Services, Data Science and Big Data Analytics: Discovering, Analyzing, Visualizing and Presenting Data, Wiley, 2015

Data Evolution and the Rise of Big Data Sources



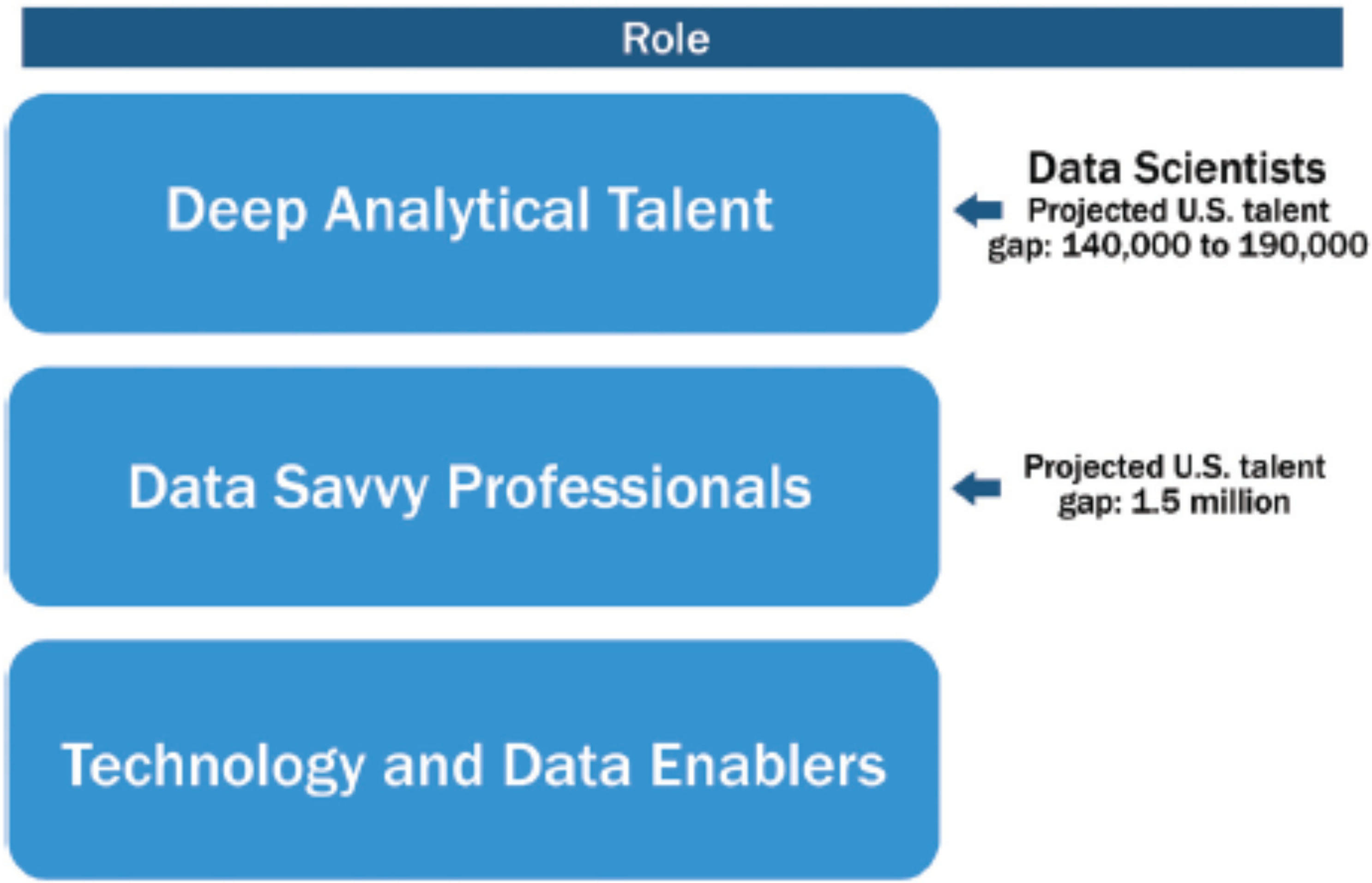
Source: EMC Education Services, Data Science and Big Data Analytics: Discovering, Analyzing, Visualizing and Presenting Data, Wiley, 2015

Emerging Big Data Ecosystem



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

Key Roles for the New Big Data Ecosystem



Note: Figures above reflect a projected talent gap in US in 2018, as shown in McKinsey May 2011 article "Big Data: The Next Frontier for Innovation, Competition, and Productivity"

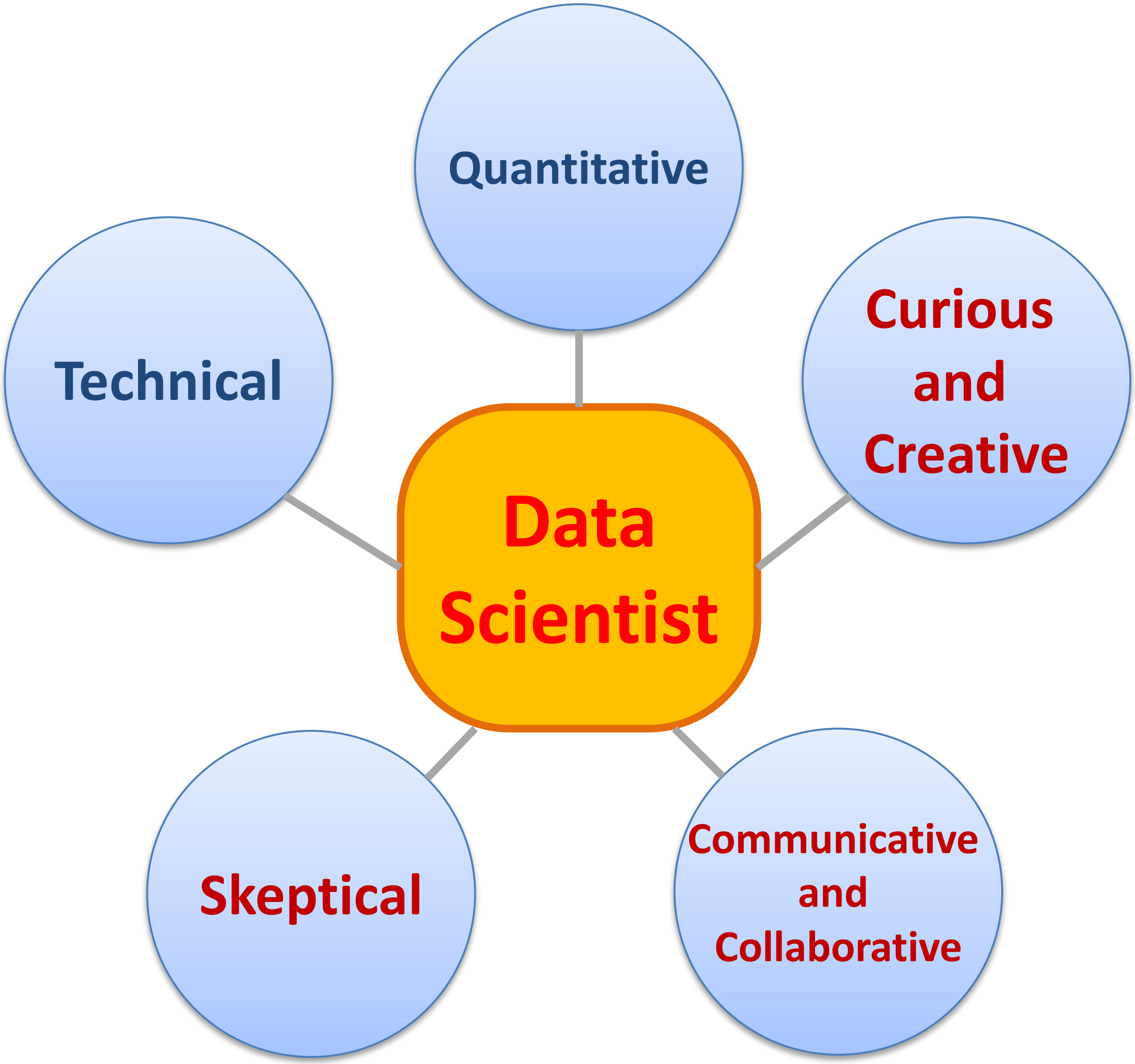
Source: EMC Education Services, Data Science and Big Data Analytics: Discovering, Analyzing, Visualizing and Presenting Data, Wiley, 2015

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

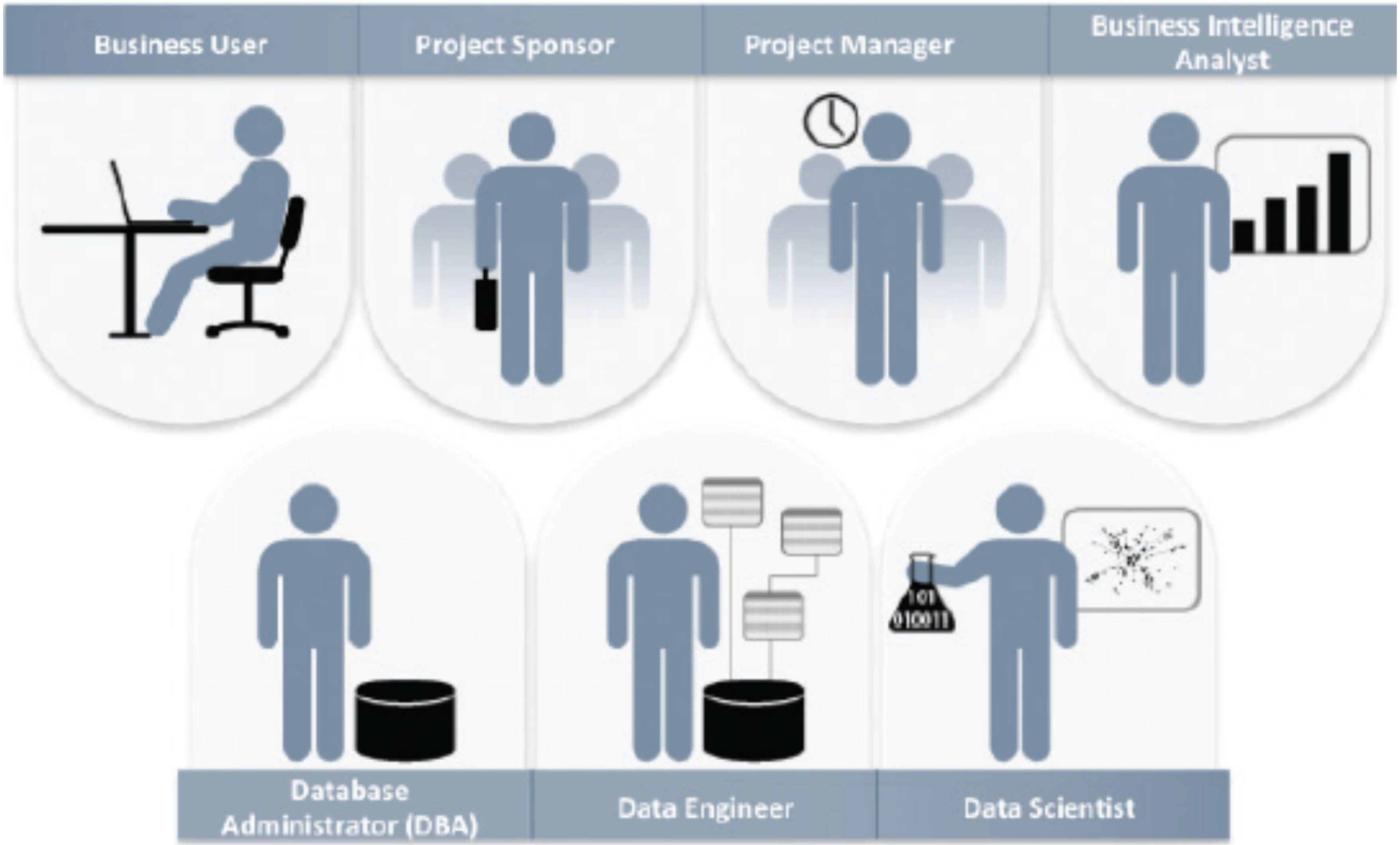
Data Scientist Profile



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

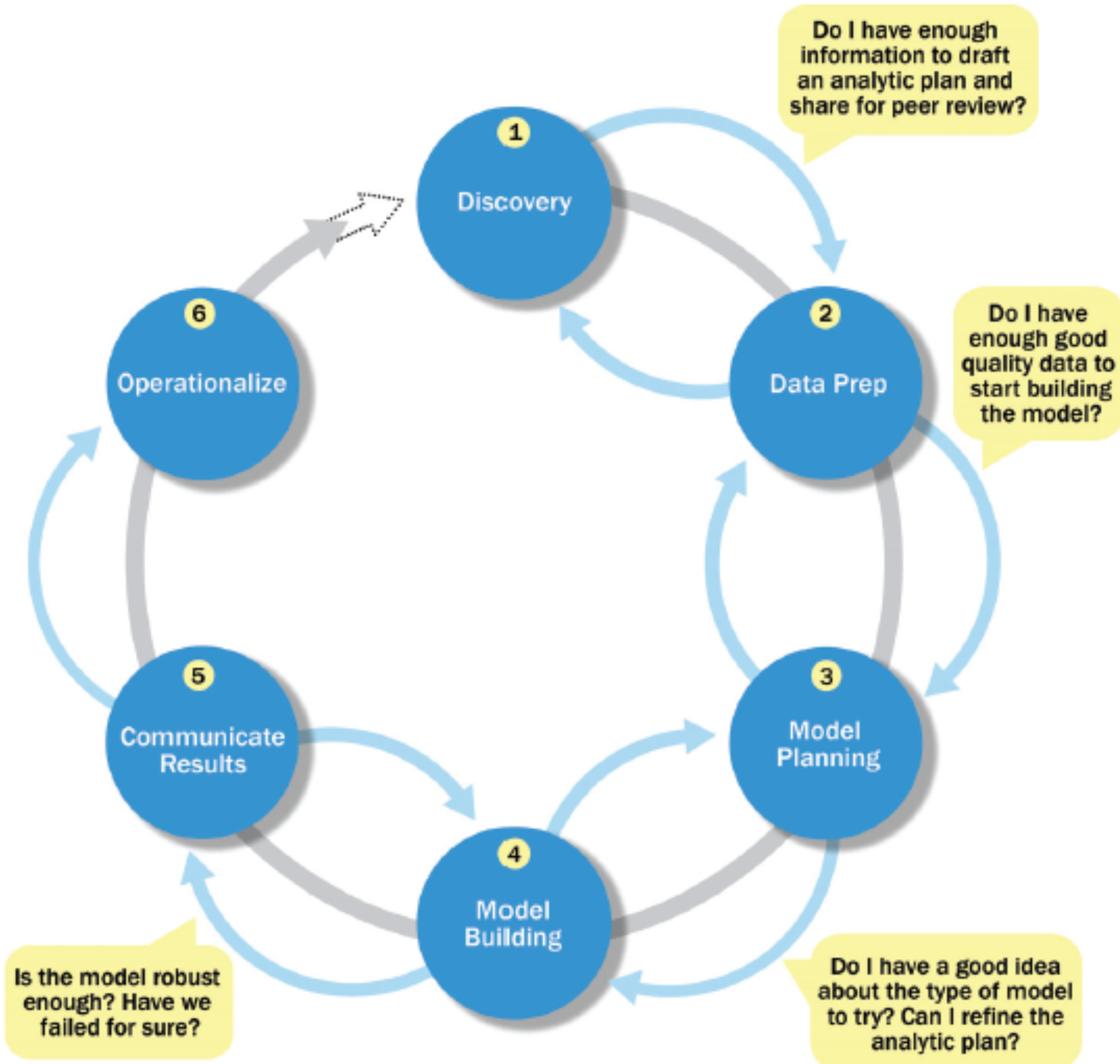
Big Data Analytics Lifecycle

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



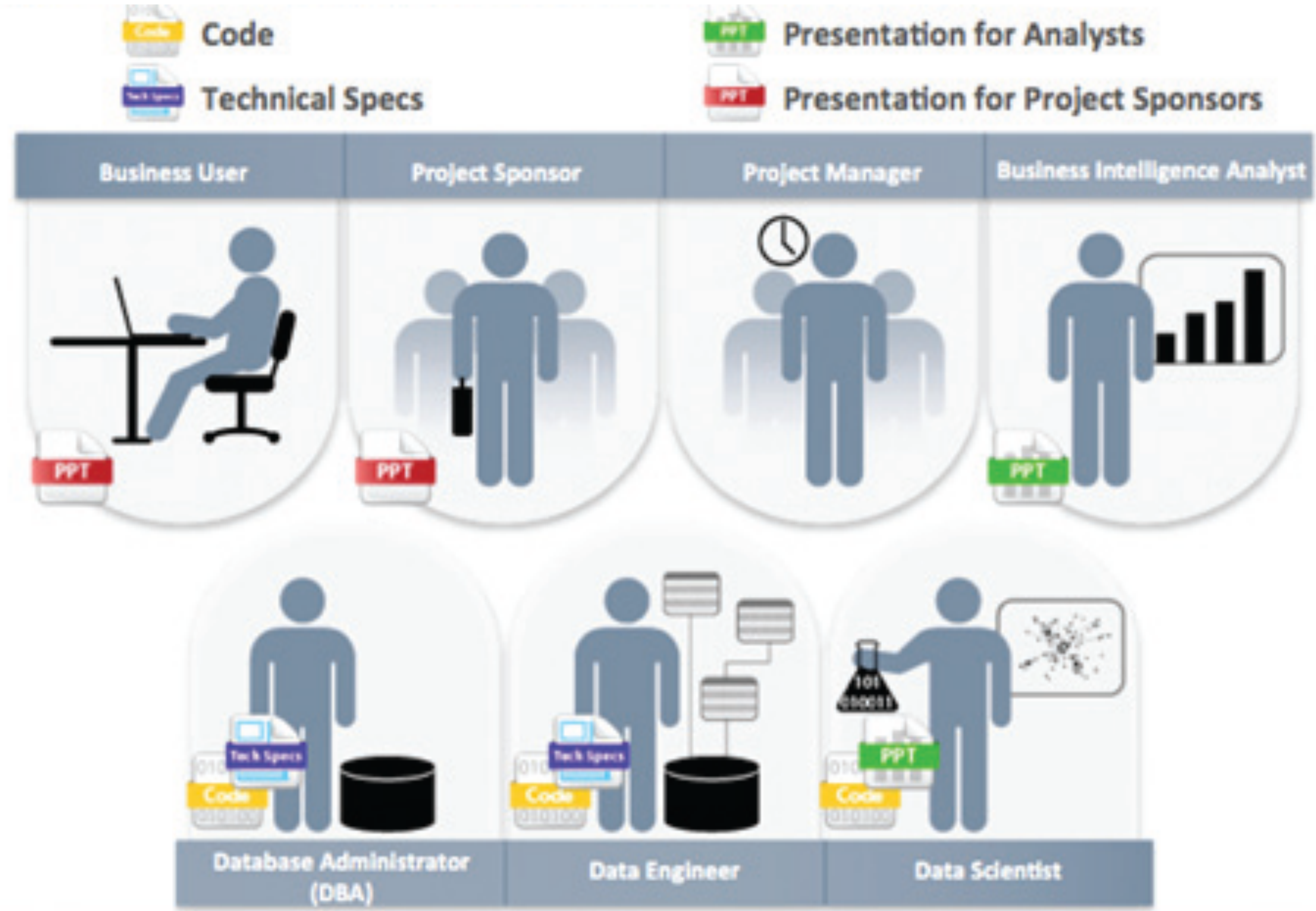
Source: EMC Education Services, Data Science and Big Data Analytics: Discovering, Analyzing, Visualizing and Presenting Data, Wiley, 2015

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)

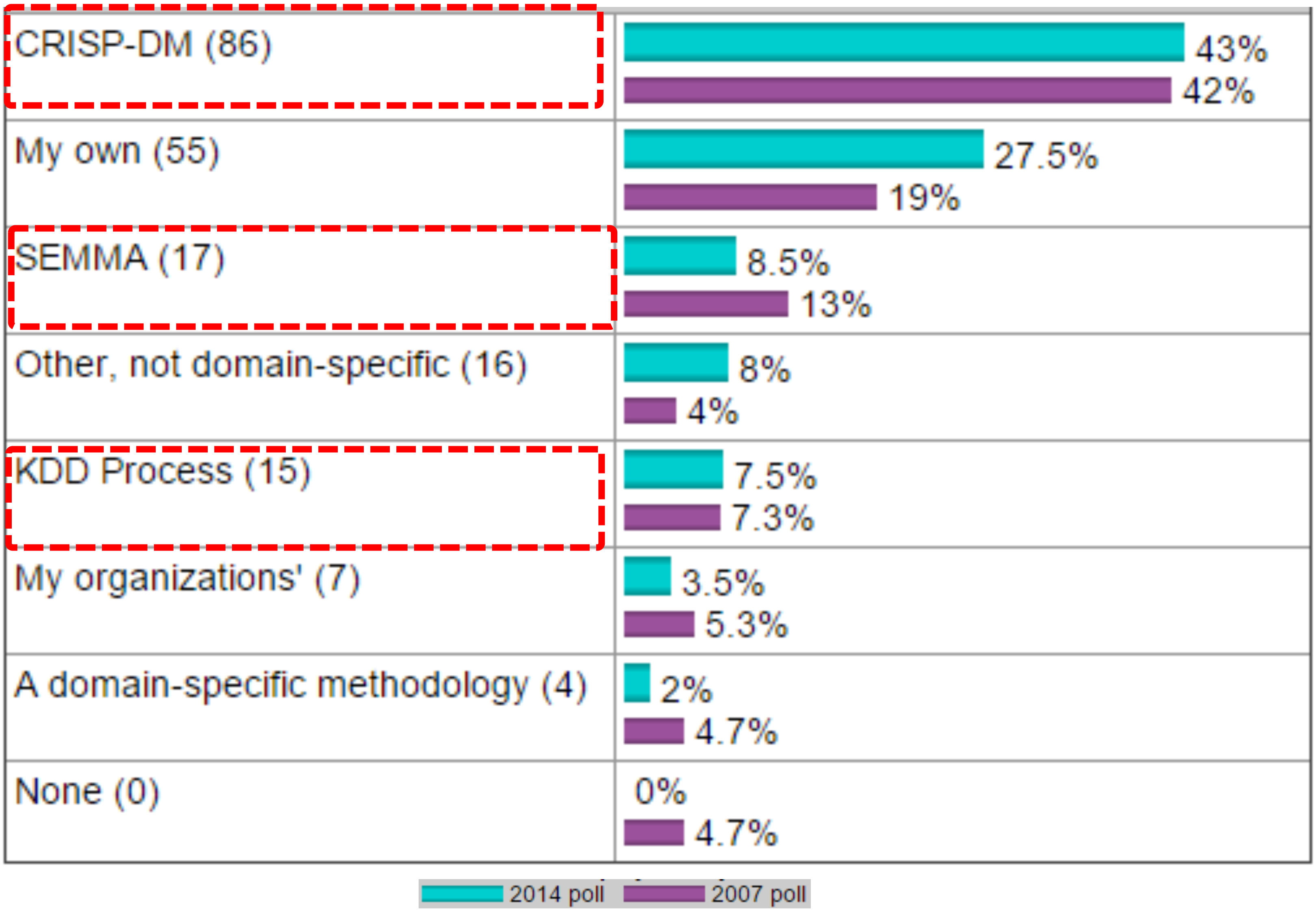
Source: Turban et al. (2011), Decision Support and Business Intelligence Systems

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.

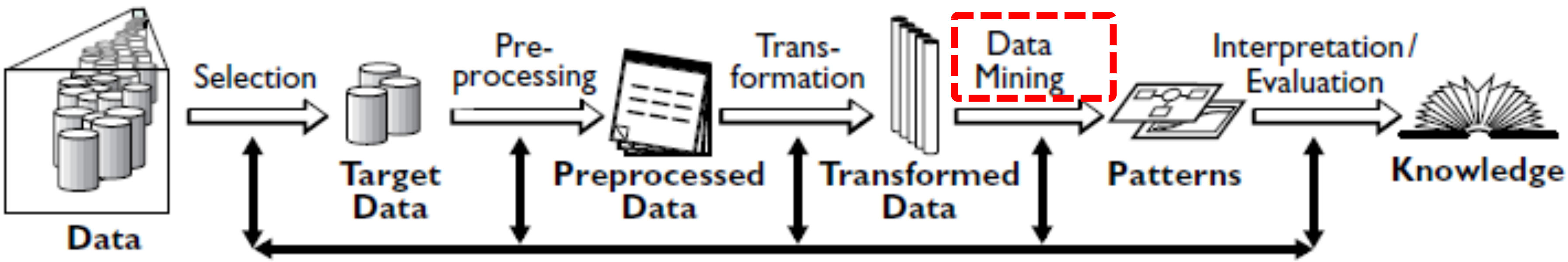
Current hardware and database technology allow efficient and inexpensive reliable data storage and access. However, whether the context is business, medicine, science, or government, the datasets themselves (in raw form) are of little direct value. What is of value is the knowledge that can be inferred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

TERRY WIDENER

Data Mining

Knowledge Discovery in Databases (KDD) Process

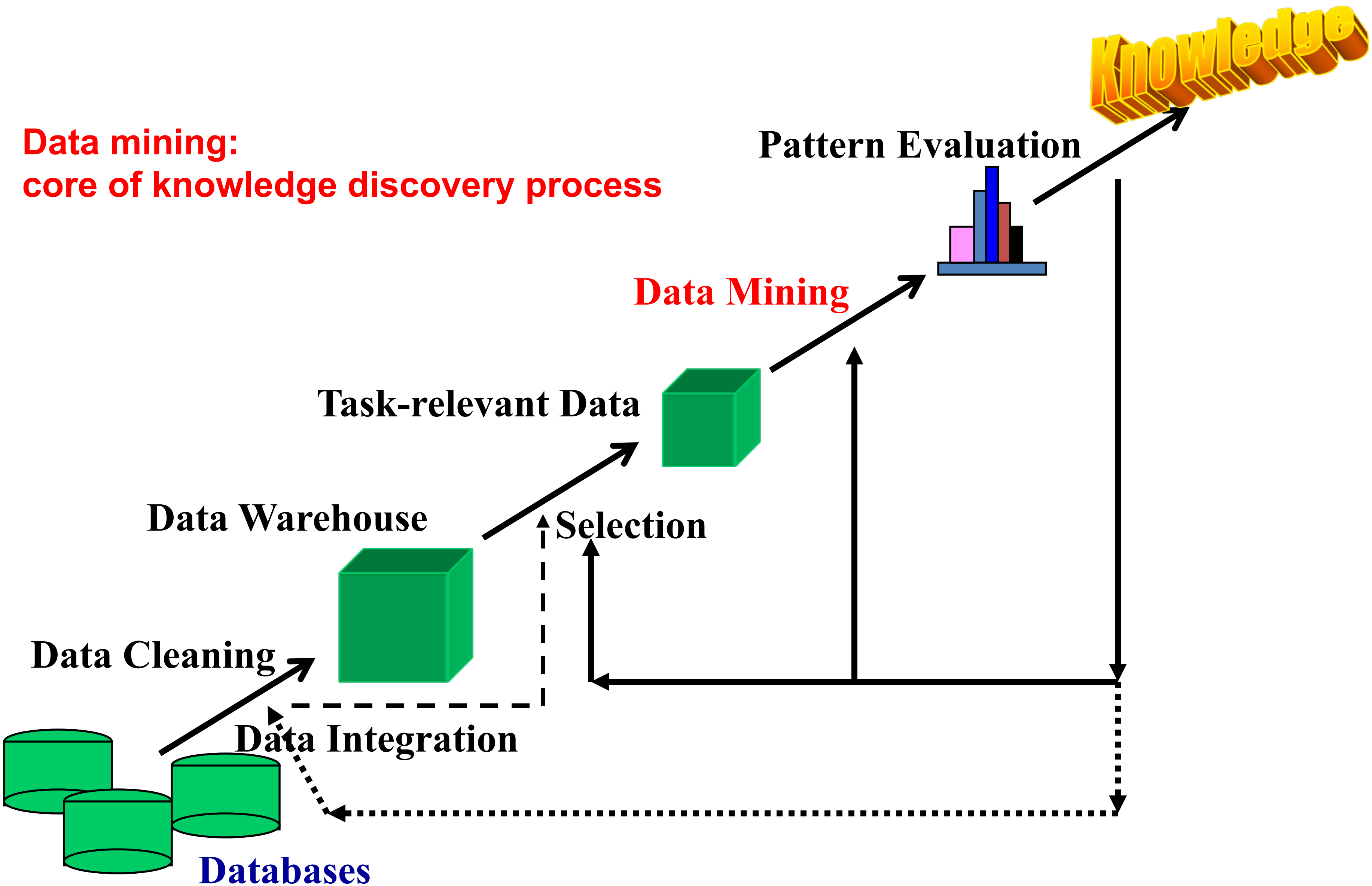
(Fayyad et al., 1996)



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.

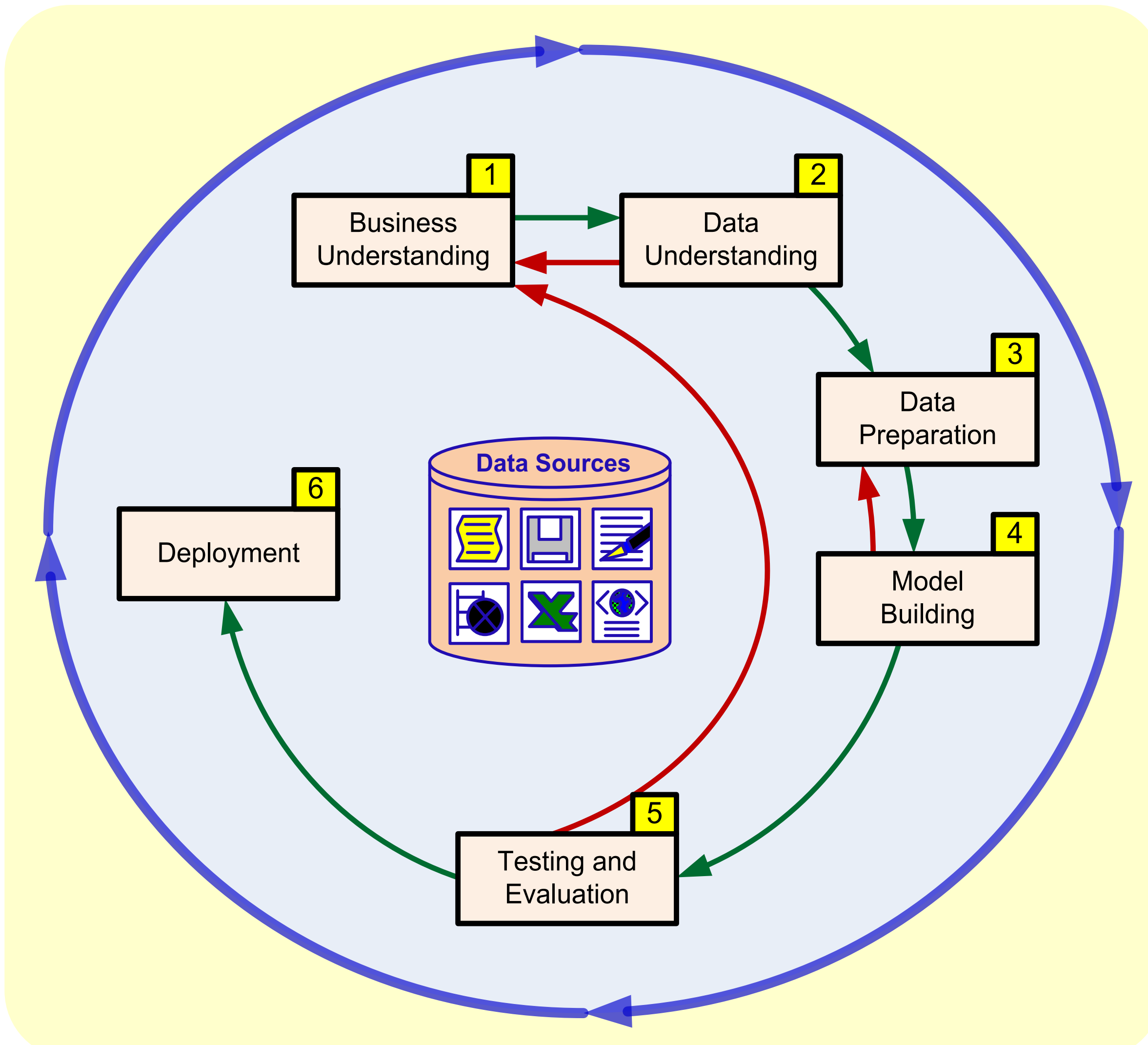
Knowledge Discovery (KDD) Process

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

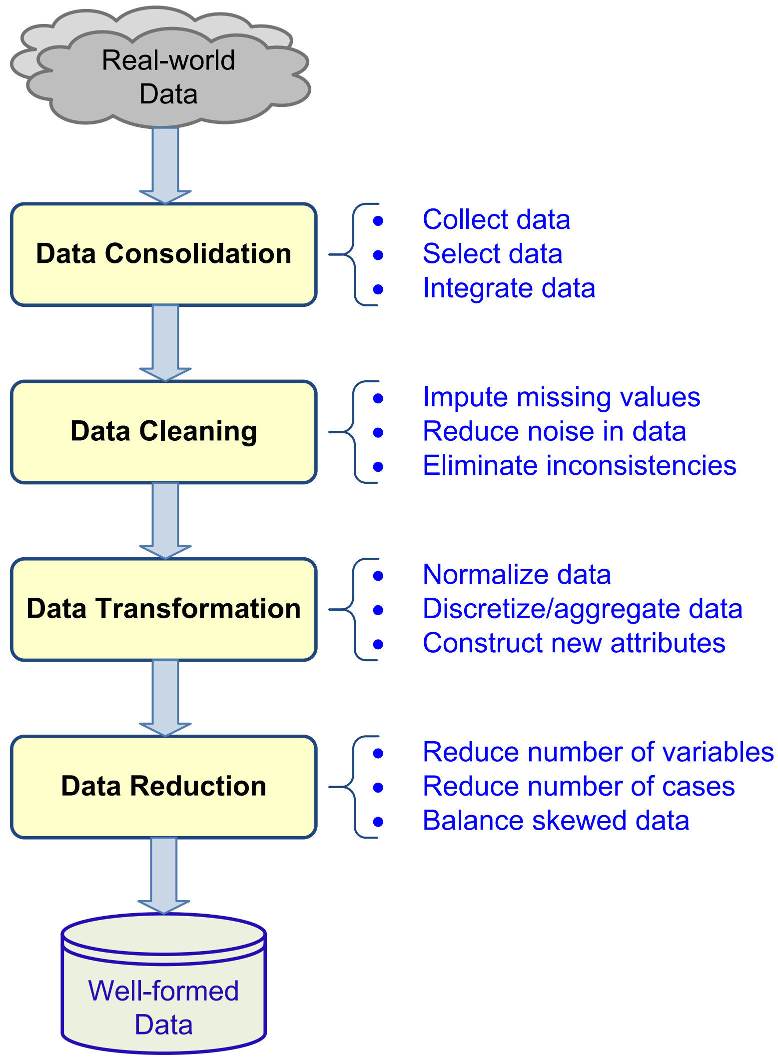
Step 4: Model Building

Step 5: Testing and Evaluation

Step 6: Deployment

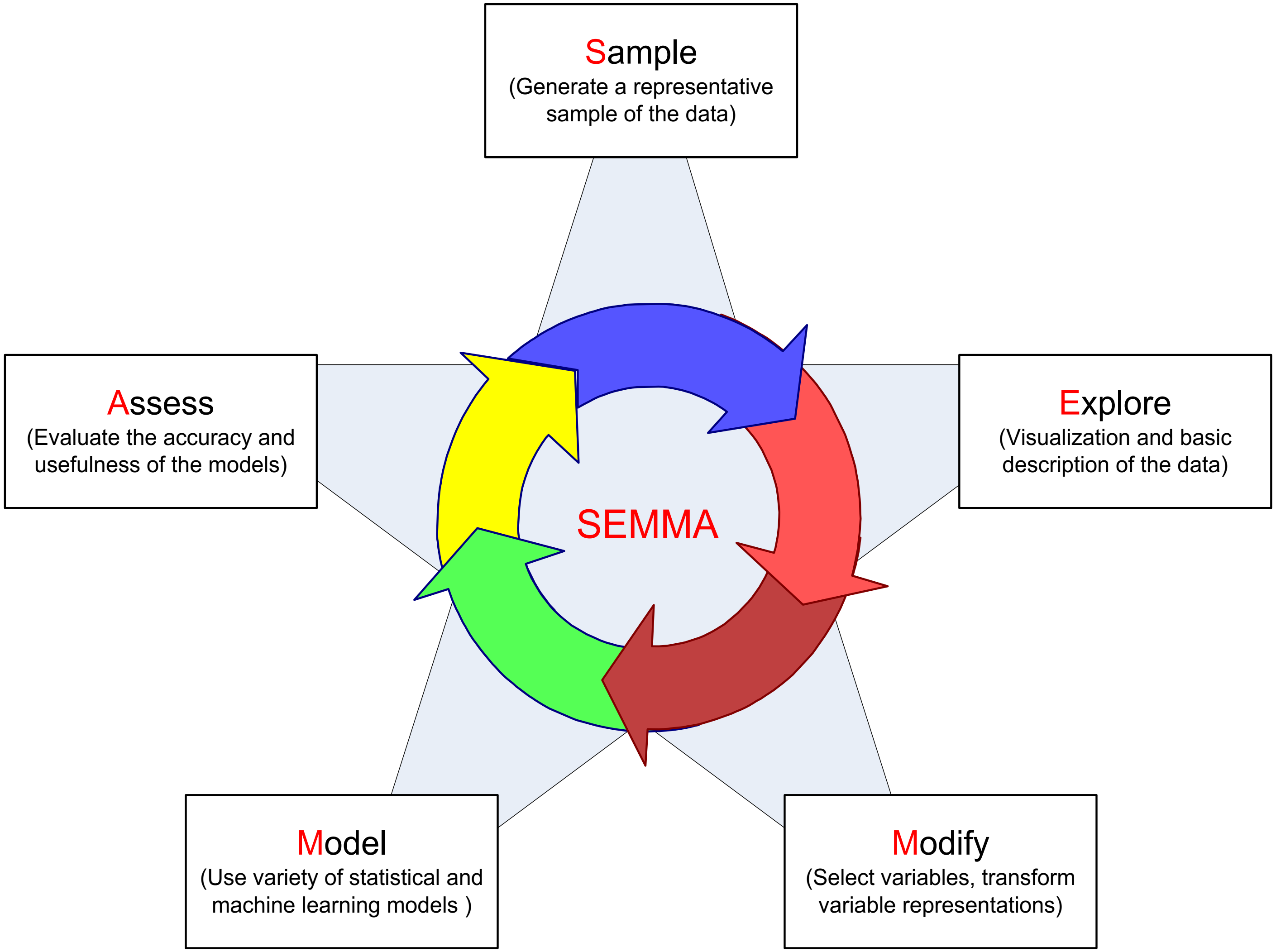
- The process is highly repetitive and experimental (DM: art versus science?)

Data Preparation – A Critical DM Task



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

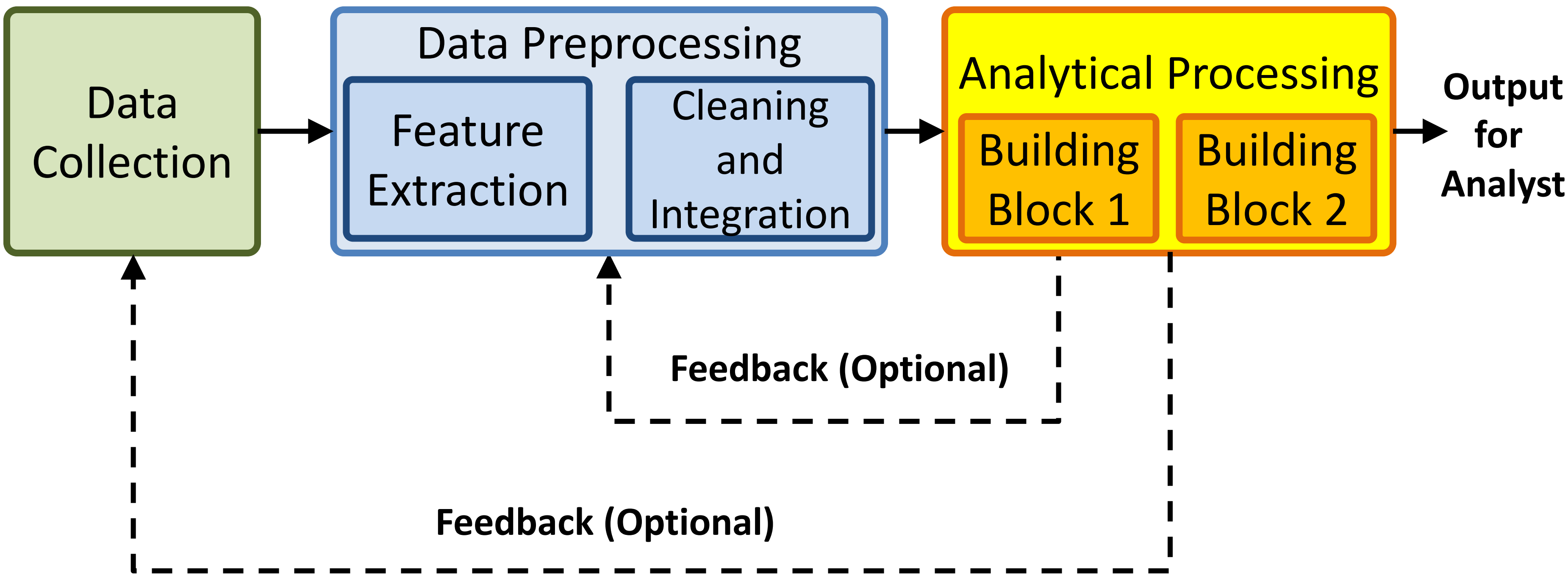
Data Mining Process: SEMMA



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

Data Mining Processing Pipeline

(Charu Aggarwal, 2015)

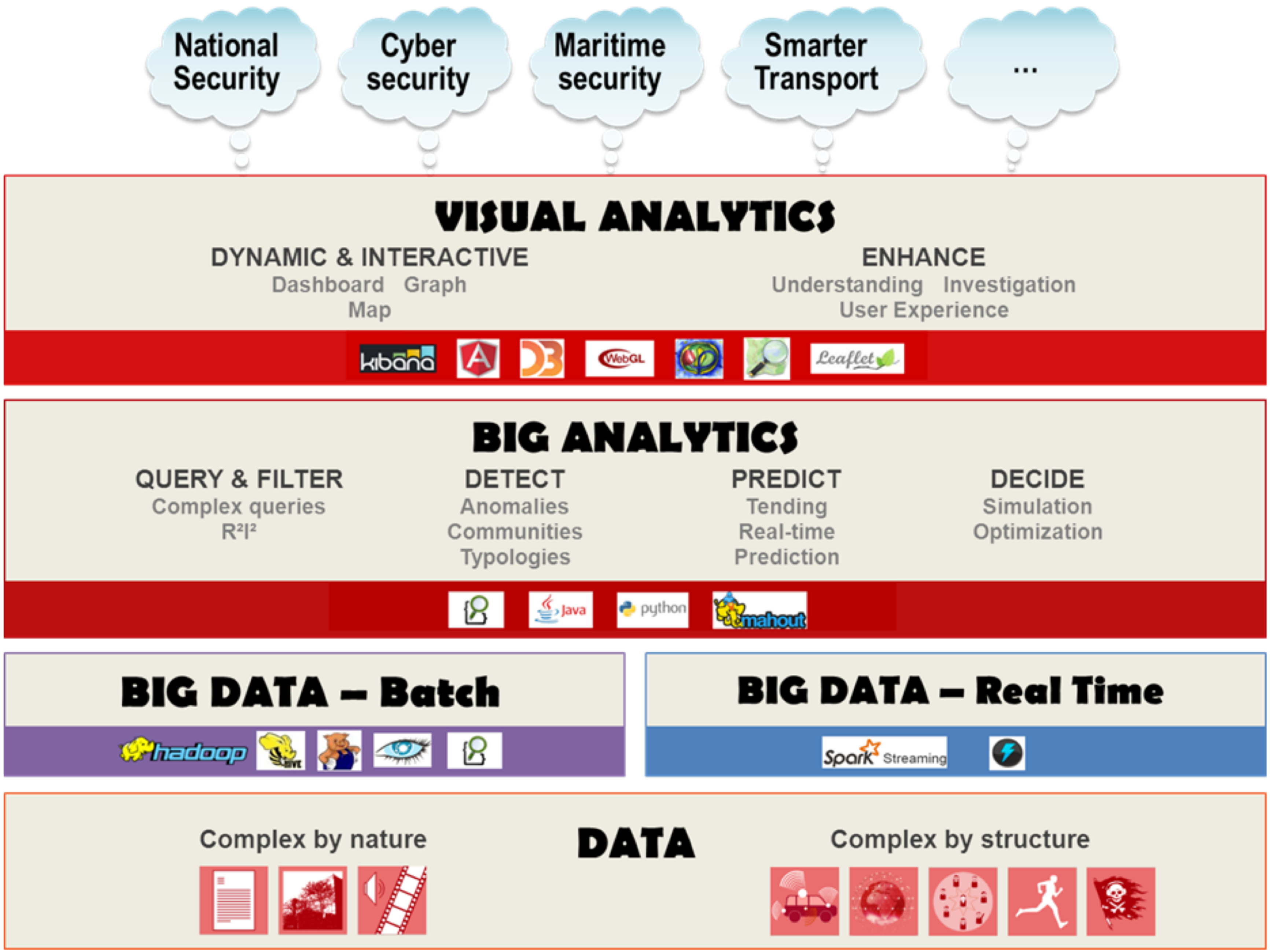


A Taxonomy for Data Mining Tasks

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

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

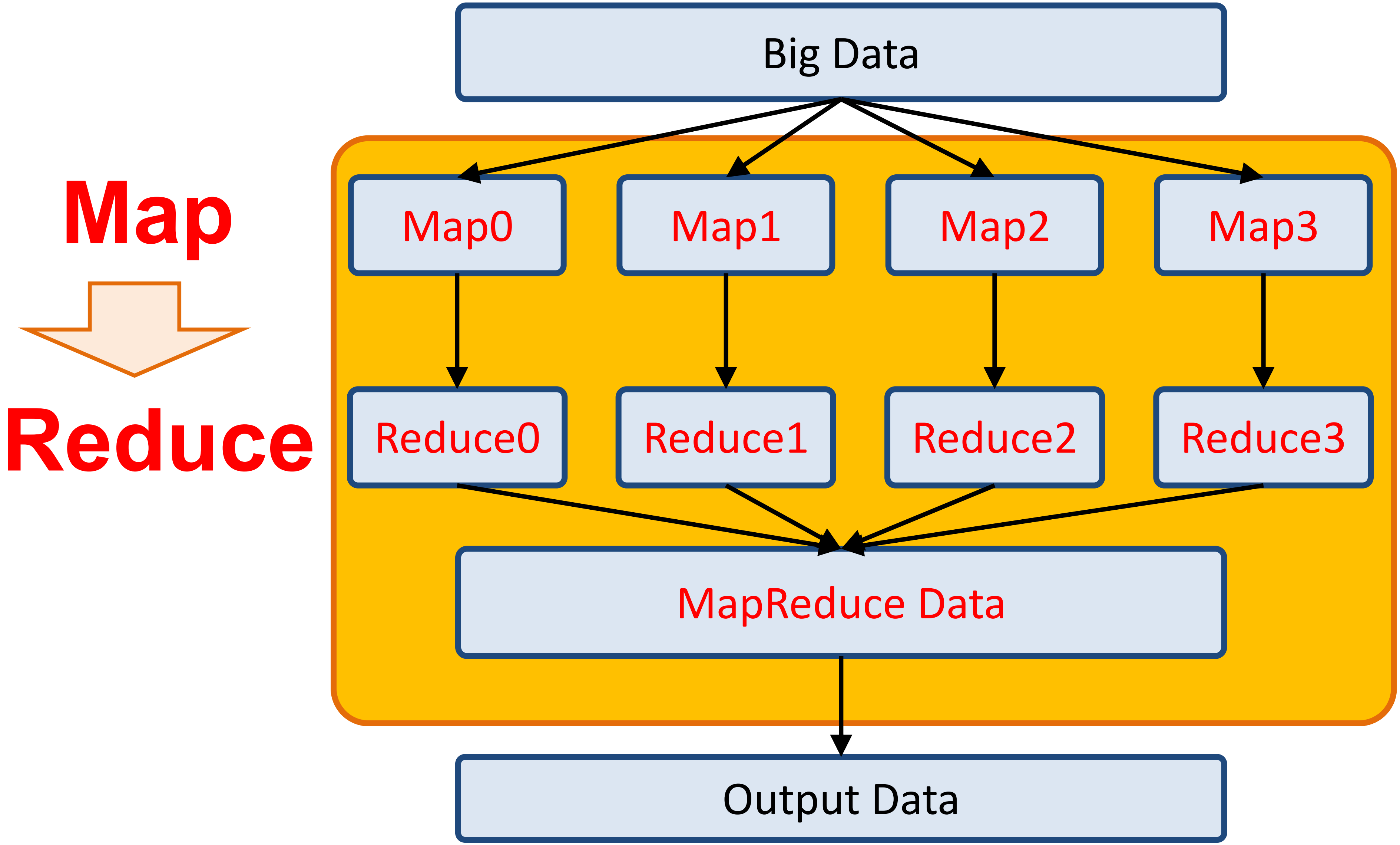
Fundamental Big Data: MapReduce Paradigm, Hadoop and Spark Ecosystem



Source: https://www.thalesgroup.com/en/worldwide/big-data/big-data-big-analytics-visual-analytics-what-does-it-all-mean

MapReduce Paradigm

MapReduce Paradigm



MapReduce Word Count

Input

Dog Love Cat
Bird Love Bird
Dog Bird Cat

Source: https://www.edureka.co/blog/mapreduce-tutorial/

MapReduce Word Count

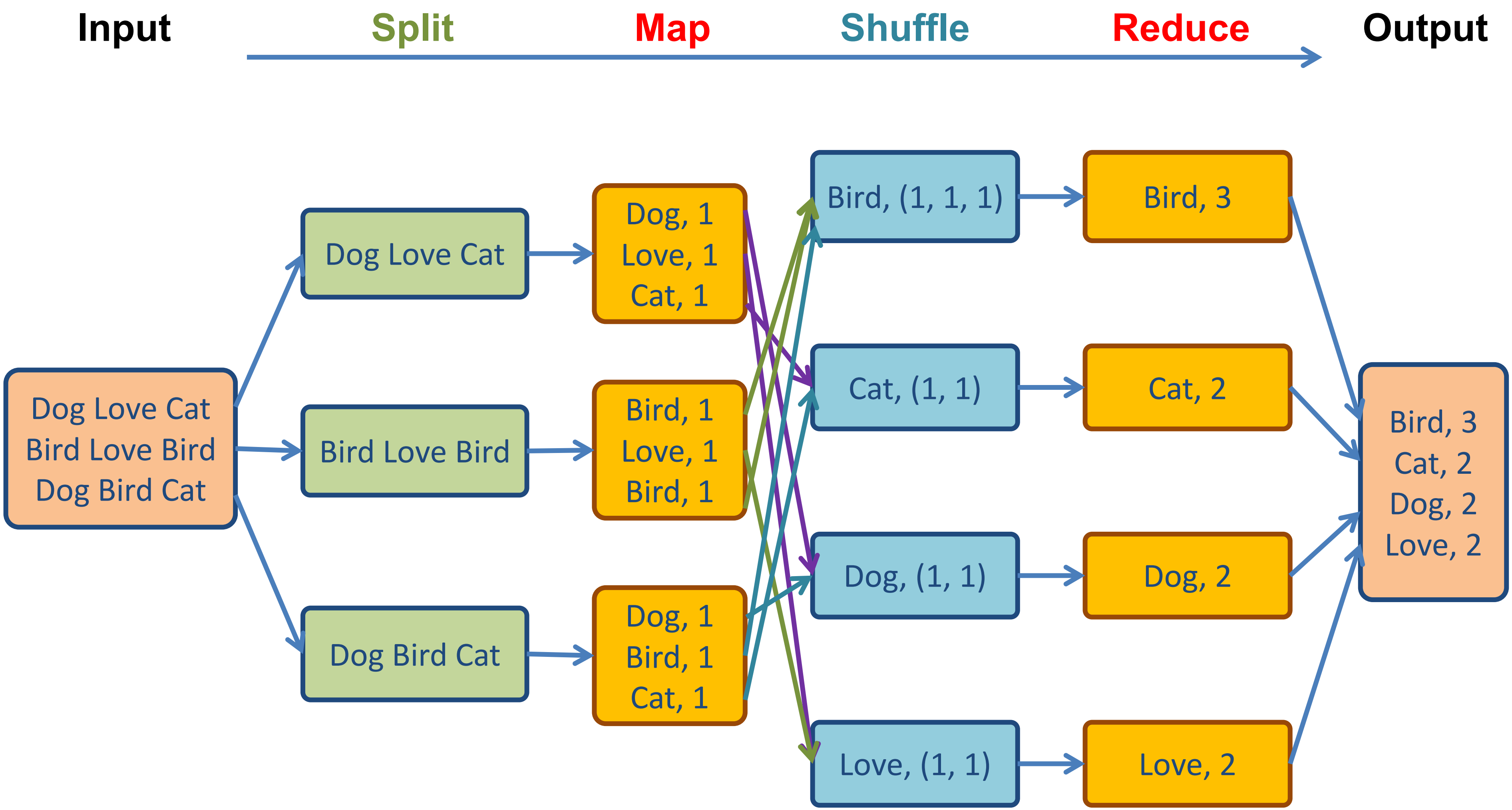
Input → **Output**

Dog Love Cat
Bird Love Bird
Dog Bird Cat

Bird, 3
Cat, 2
Dog, 2
Love, 2

Source: https://www.edureka.co/blog/mapreduce-tutorial/

MapReduce Word Count



Source: https://www.edureka.co/blog/mapreduce-tutorial/

Hadoop Ecosystem

The **Apache™ Hadoop®** project develops **open-source software** for reliable, scalable, **distributed computing**.

Source: http://hadoop.apache.org/

hadoop

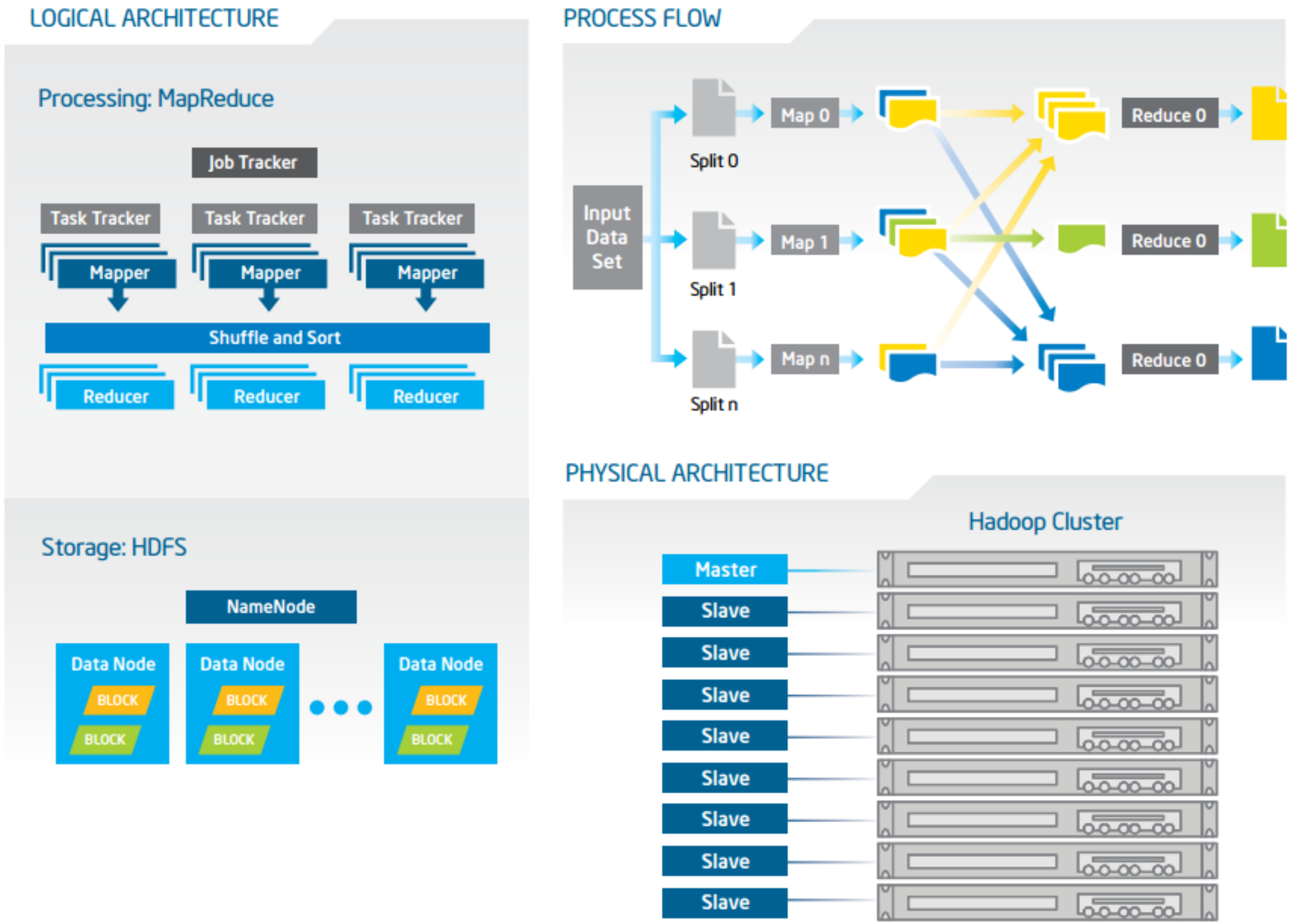
MapReduce — Processing

+

HDFS — Storage

Source: http://hadoop.apache.org/

Big Data with Hadoop Architecture

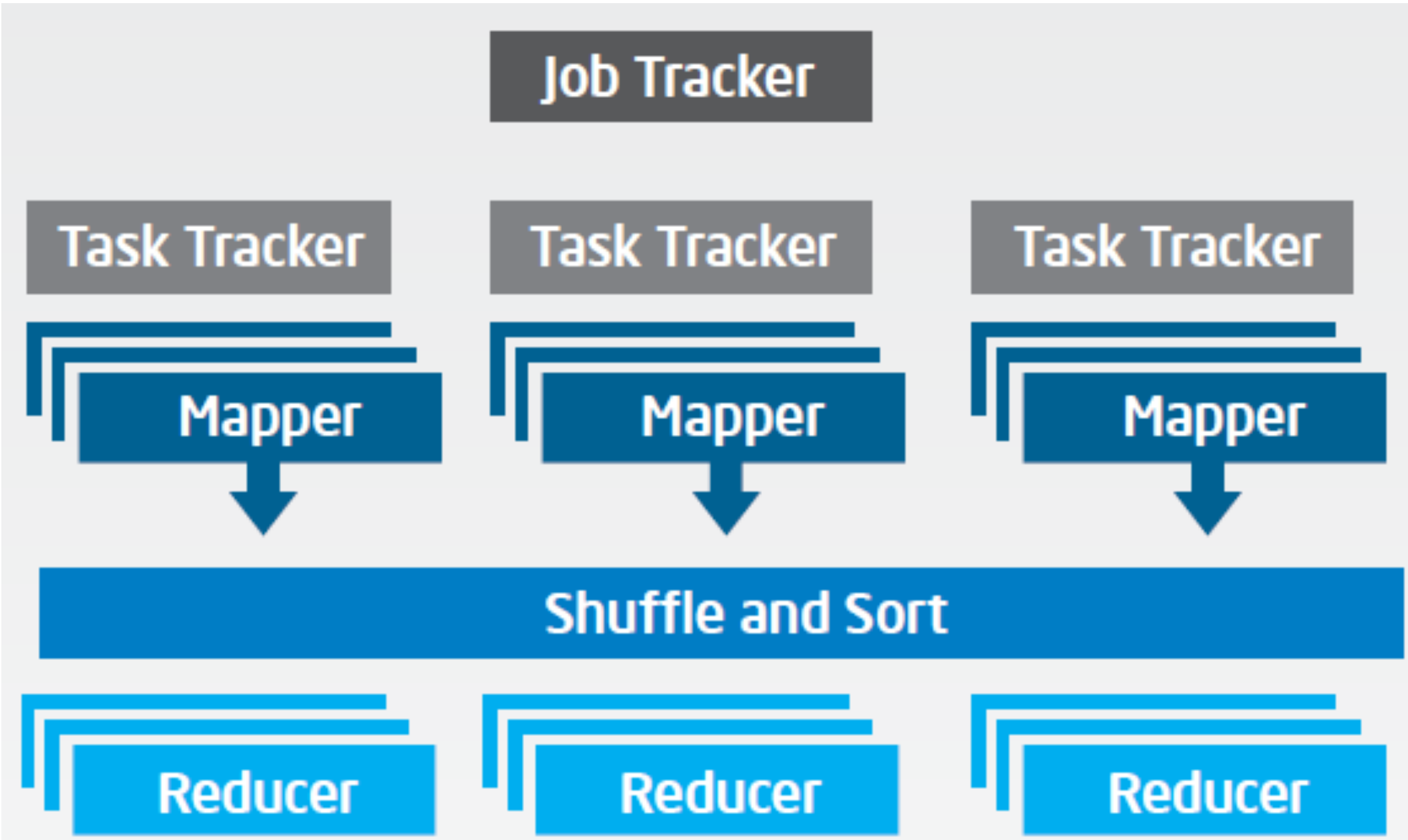


Source: https://software.intel.com/sites/default/files/article/402274/etl-big-data-with-hadoop.pdf

Big Data with Hadoop Architecture

Logical Architecture

Processing: MapReduce

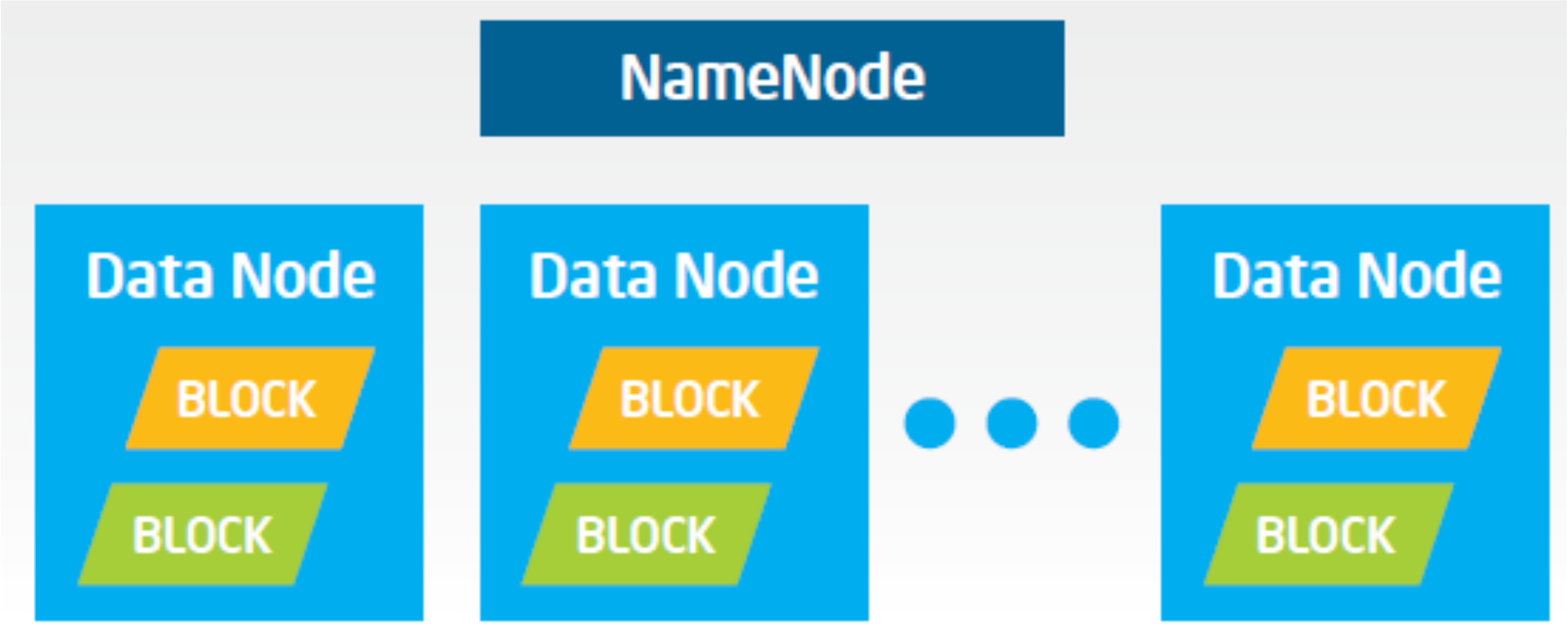


Source: https://software.intel.com/sites/default/files/article/402274/etl-big-data-with-hadoop.pdf

Big Data with Hadoop Architecture

Logical Architecture

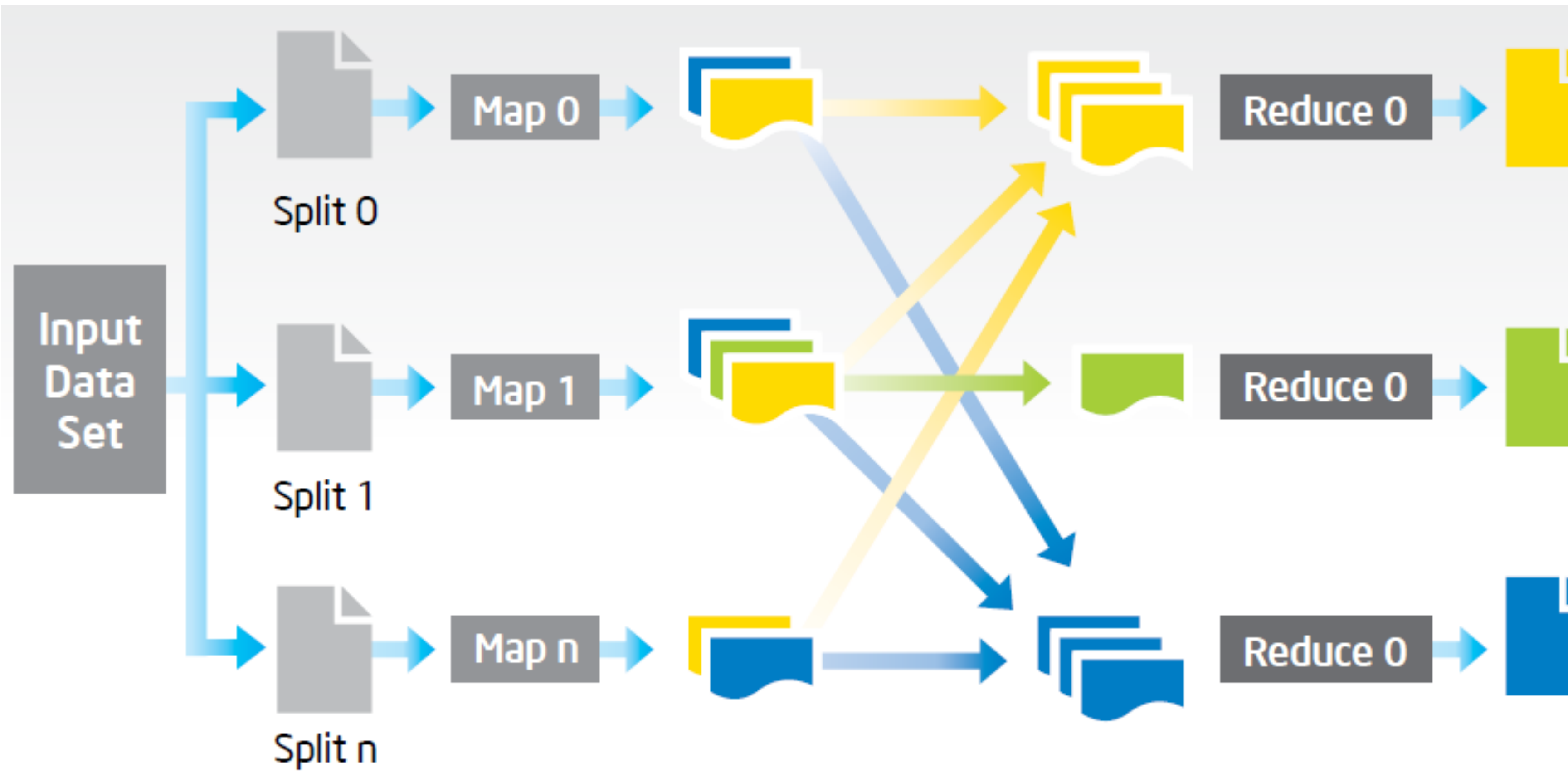
Storage: HDFS



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

Big Data with Hadoop Architecture

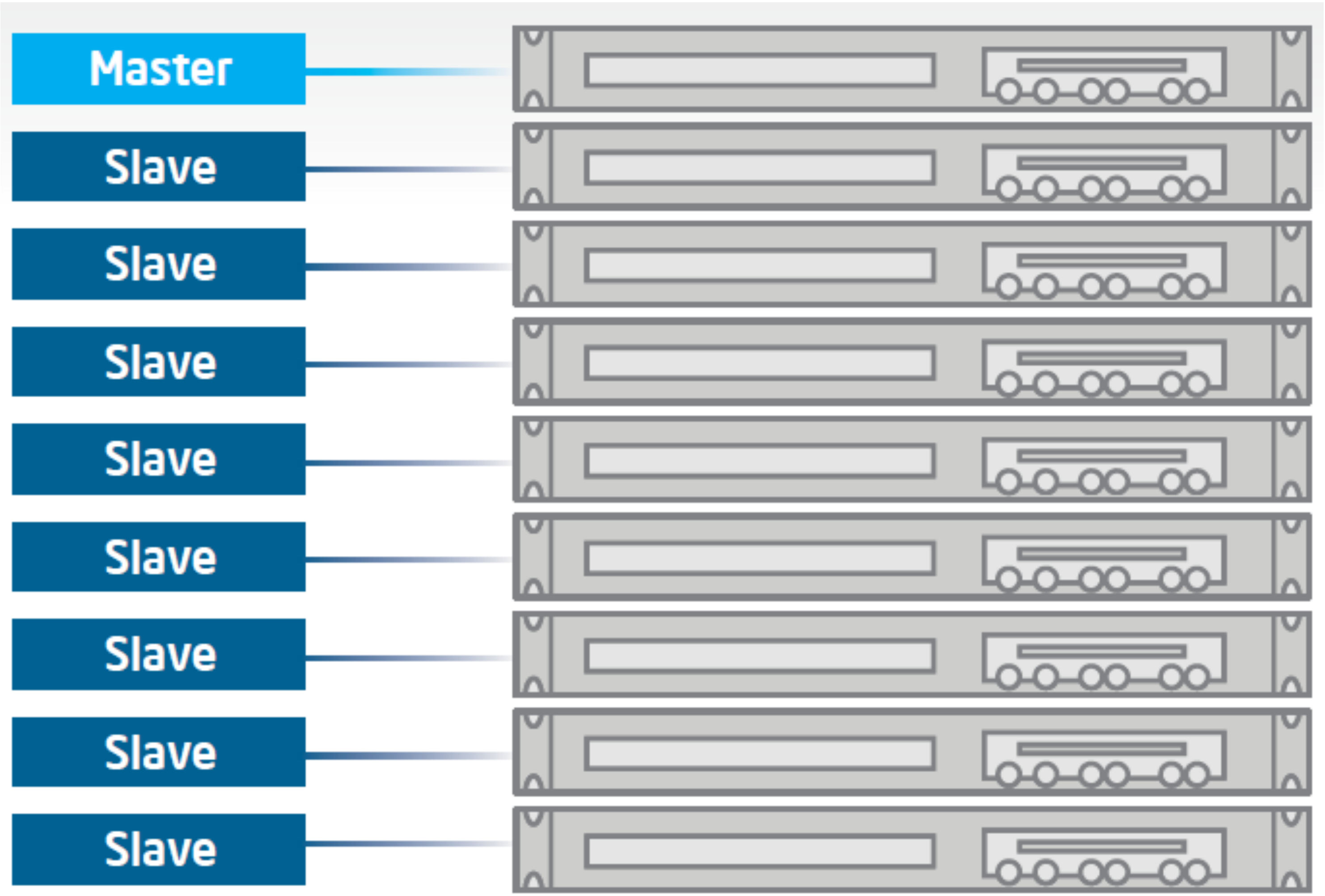
Process Flow



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

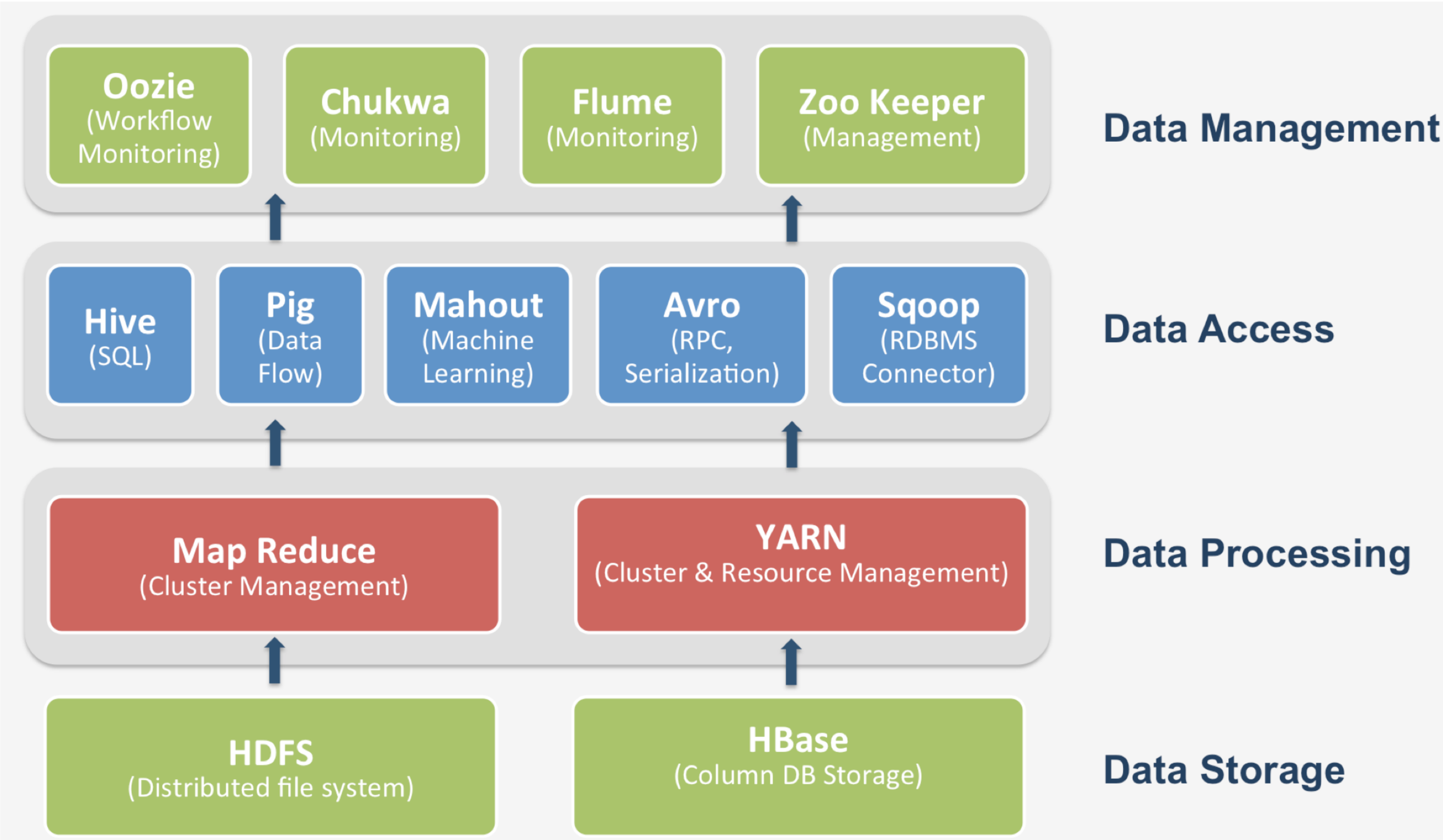
Big Data with Hadoop Architecture

Hadoop Cluster



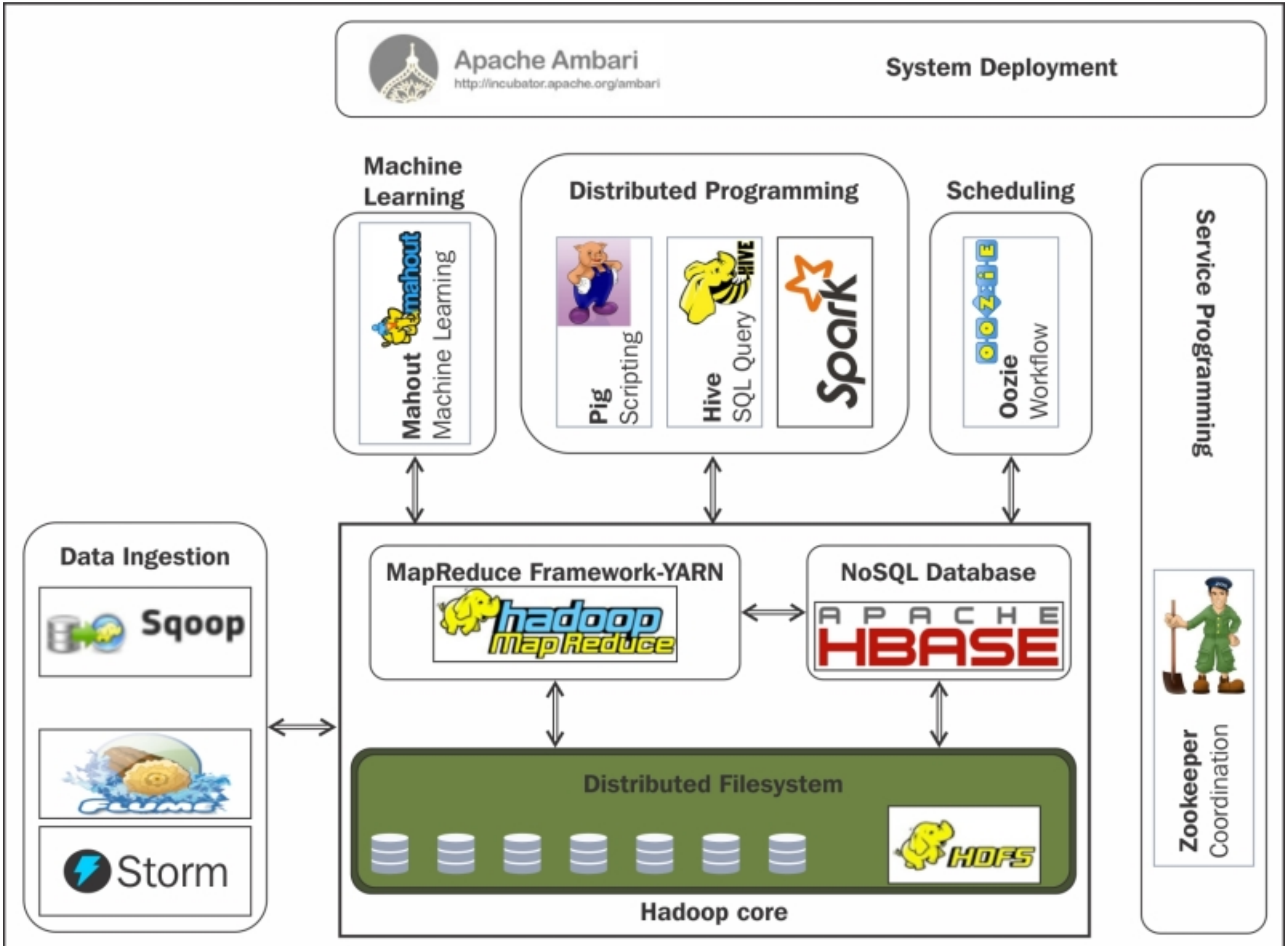
Source: https://software.intel.com/sites/default/files/article/402274/etl-big-data-with-hadoop.pdf

Hadoop Ecosystem



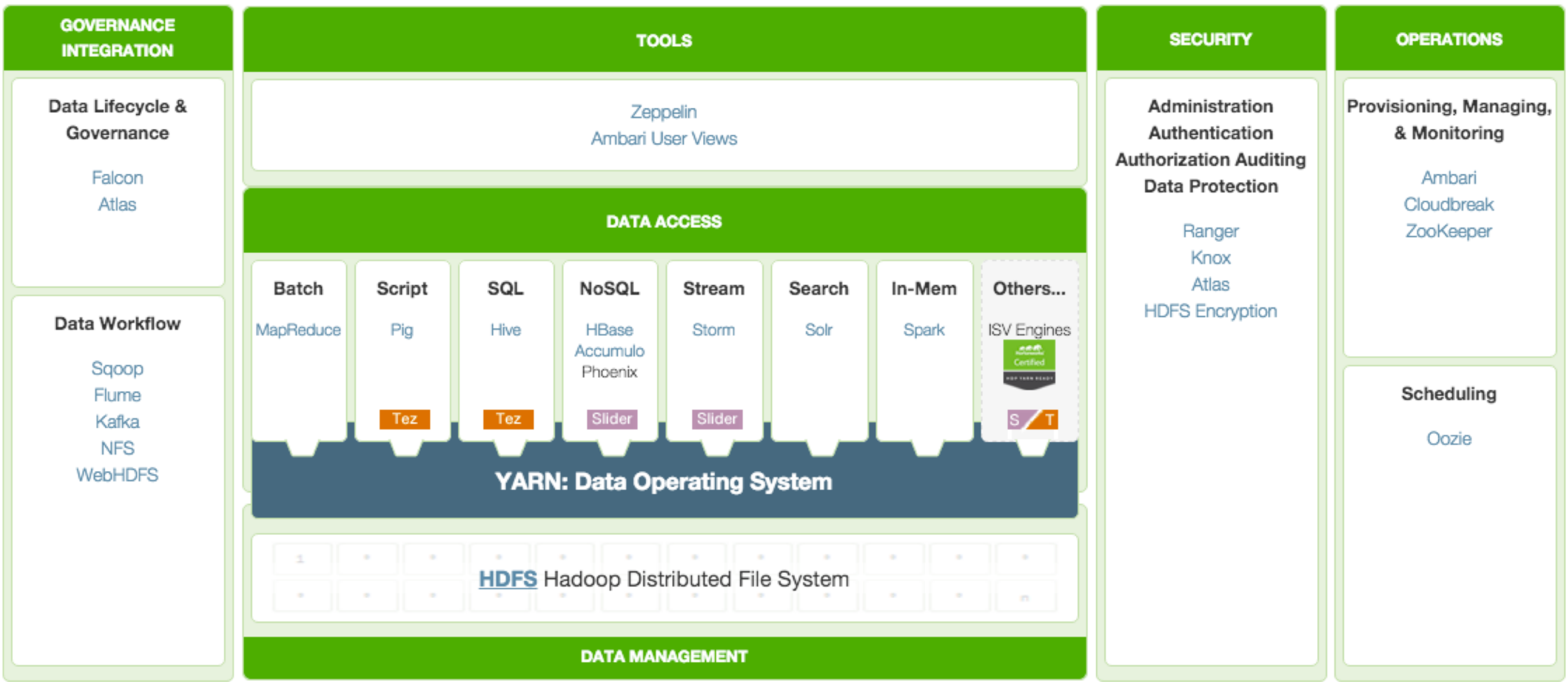
Source: https://savvycomsoftware.com/what-you-need-to-know-about-hadoop-and-its-ecosystem/

Hadoop Ecosystem



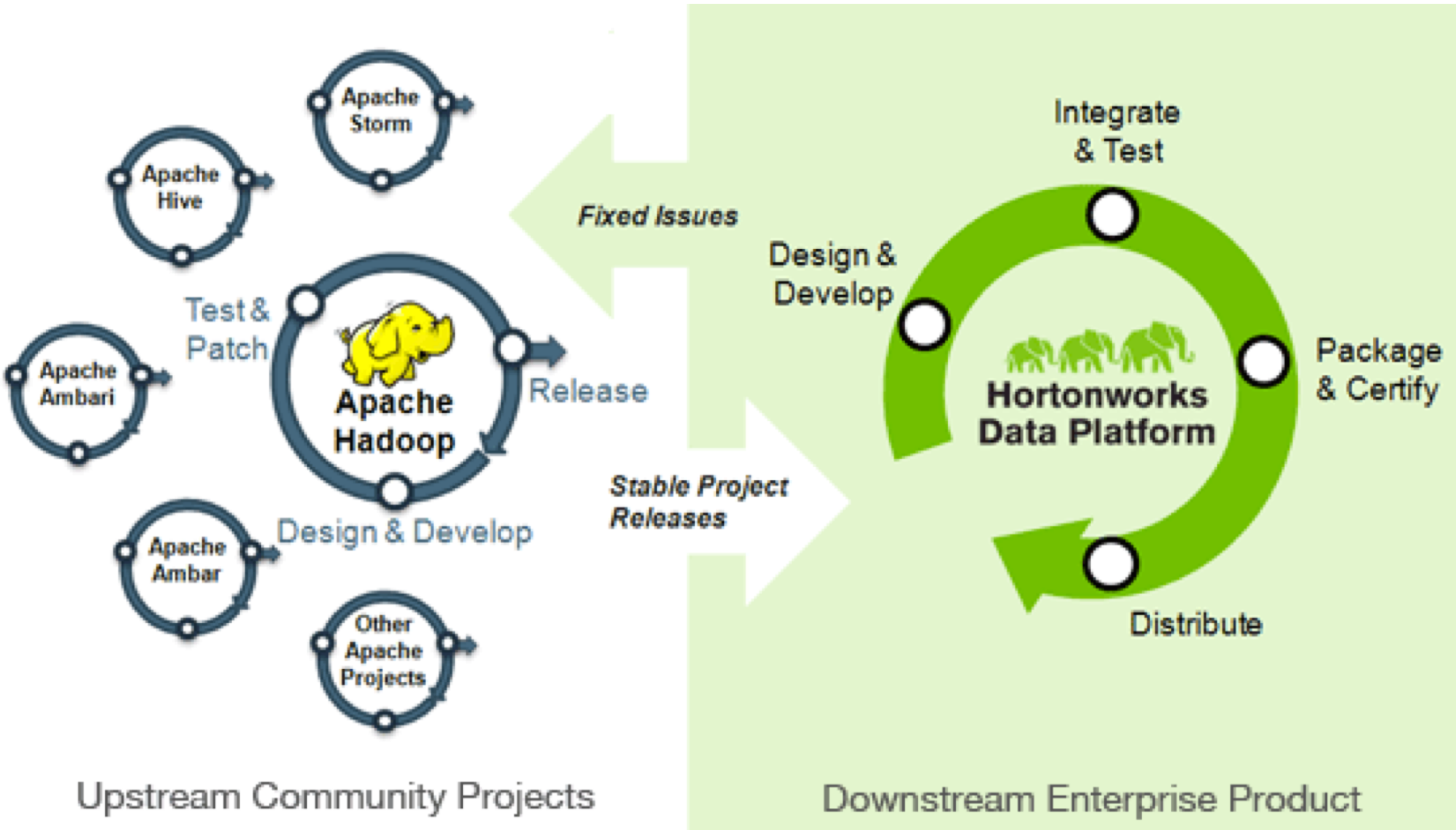
Source: Shiva Achari (2015), Hadoop Essentials - Tackling the Challenges of Big Data with Hadoop, Packt Publishing

HDP (Hortonworks Data Platform)
A Complete Enterprise Hadoop Data Platform



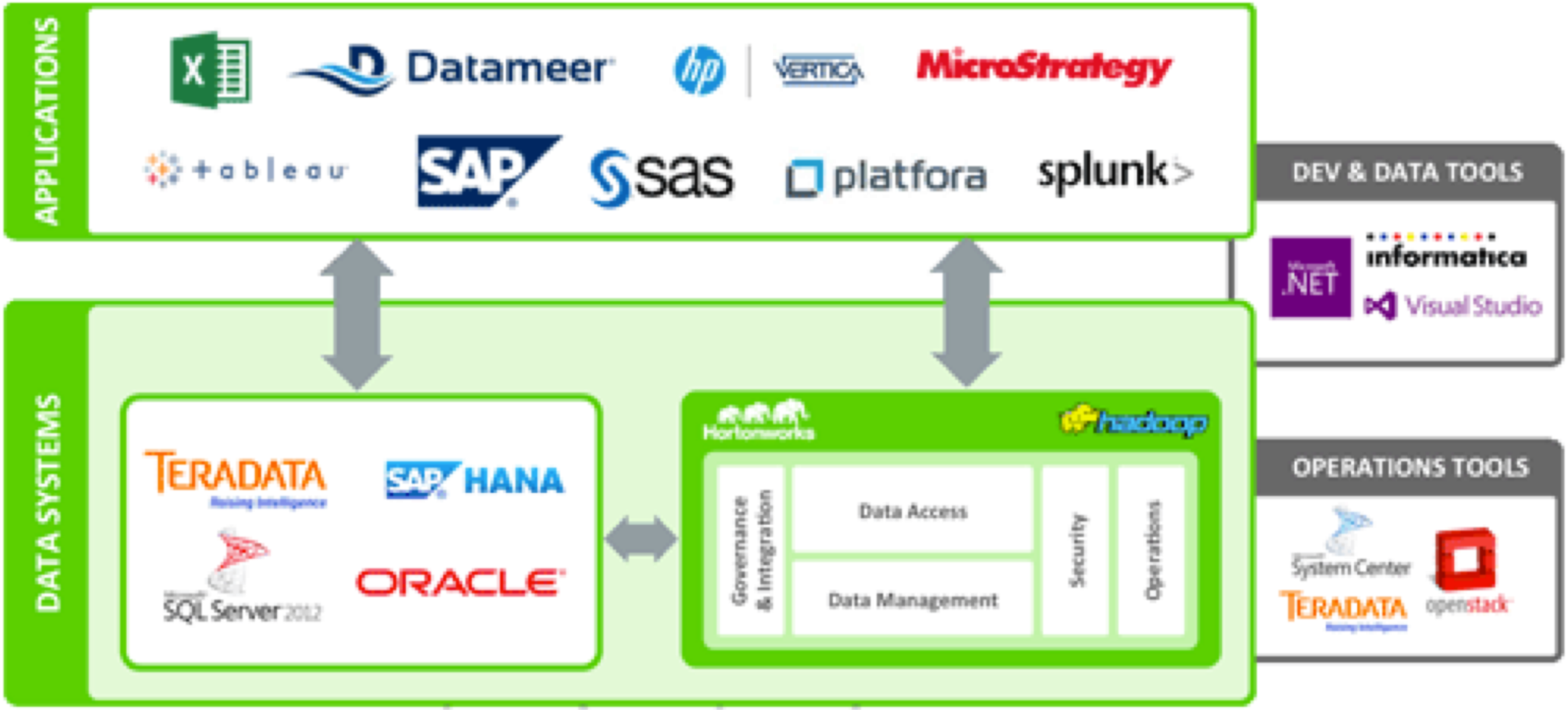
Source: http://hortonworks.com/hdp/

Apache Hadoop
Hortonworks Data Platform



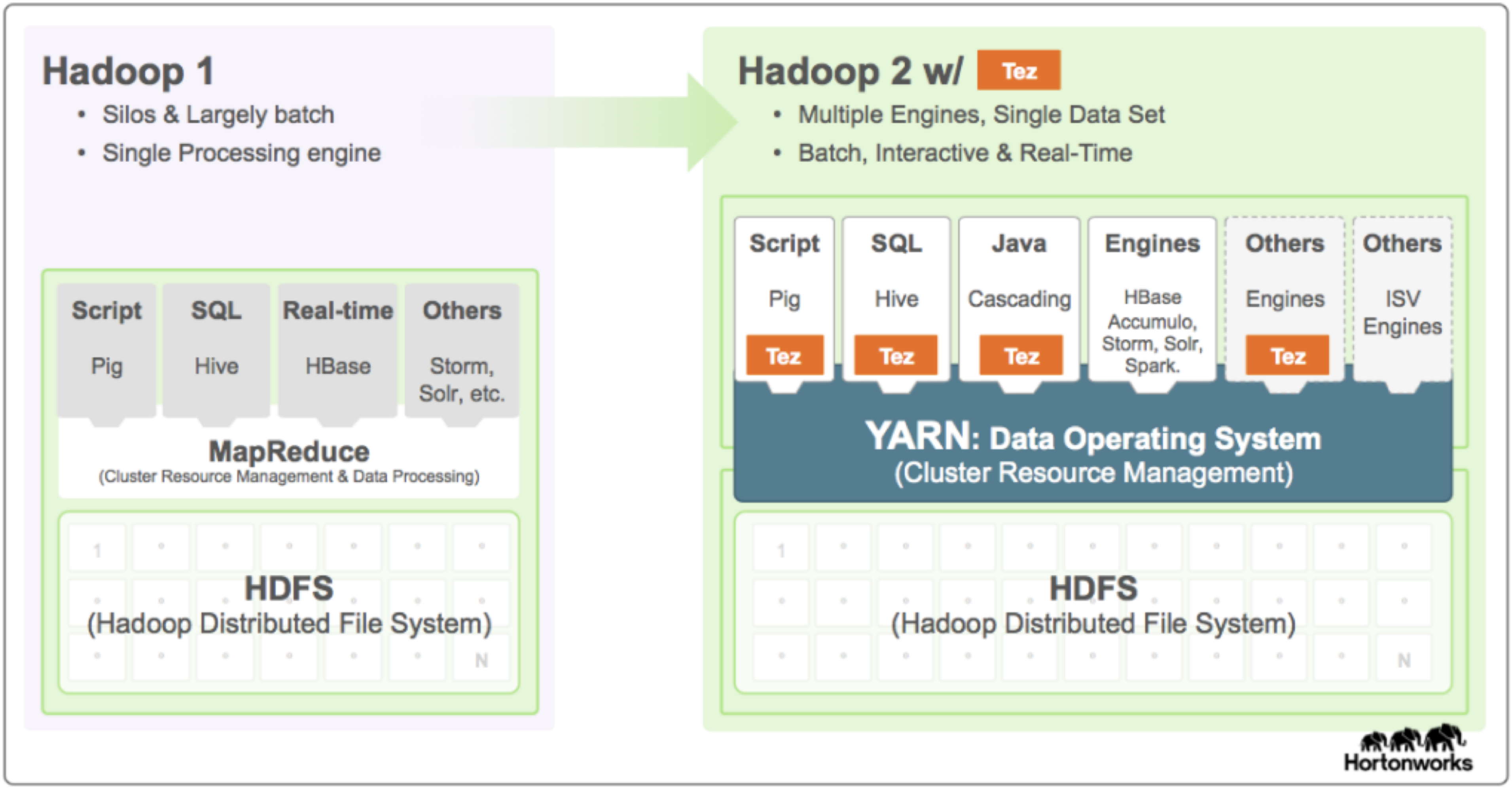
Source: http://hortonworks.com/hdp/

Hadoop and Data Analytics Tools



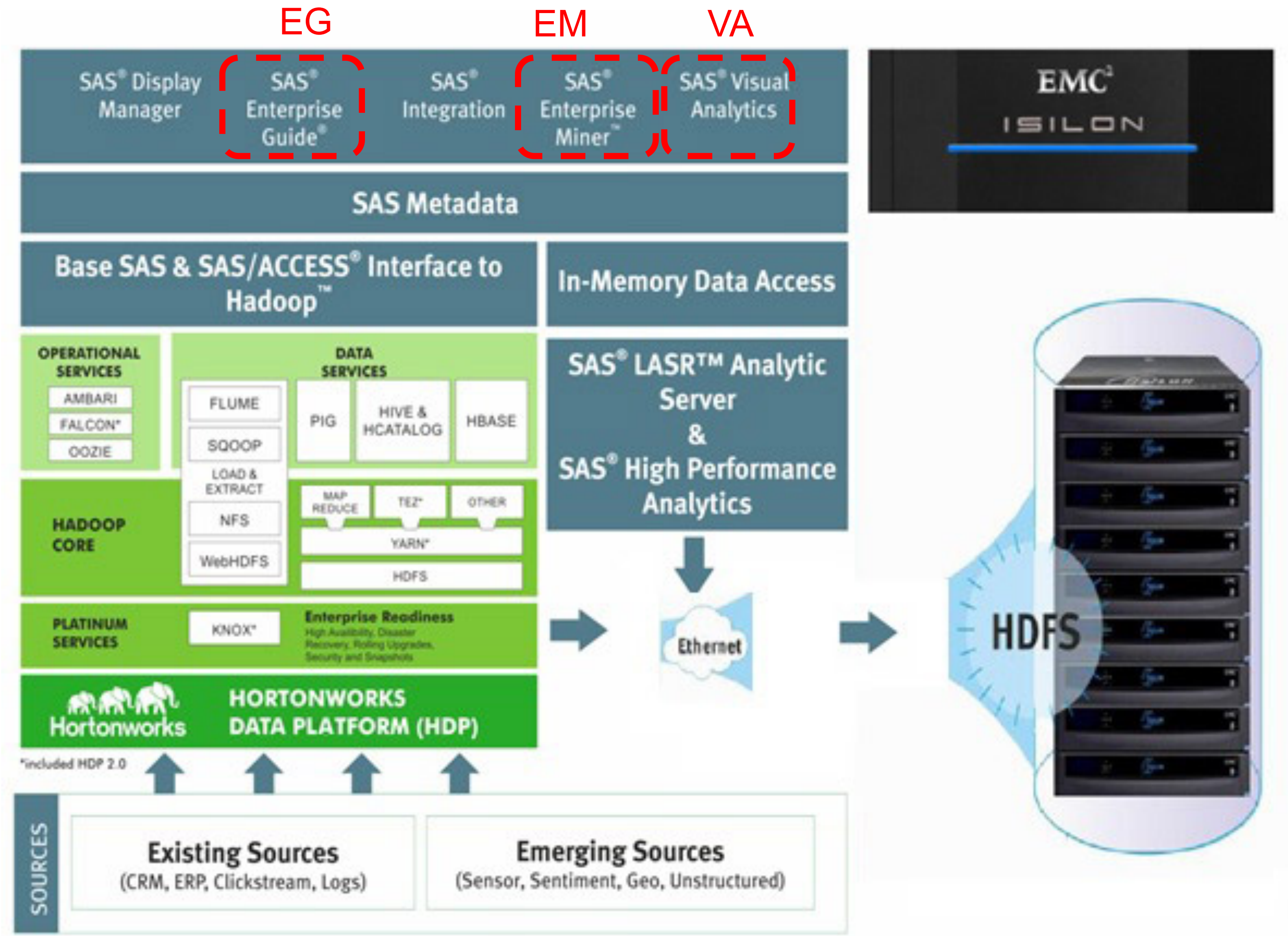
Source: http://hortonworks.com/hdp/

Hadoop 1 → Hadoop 2



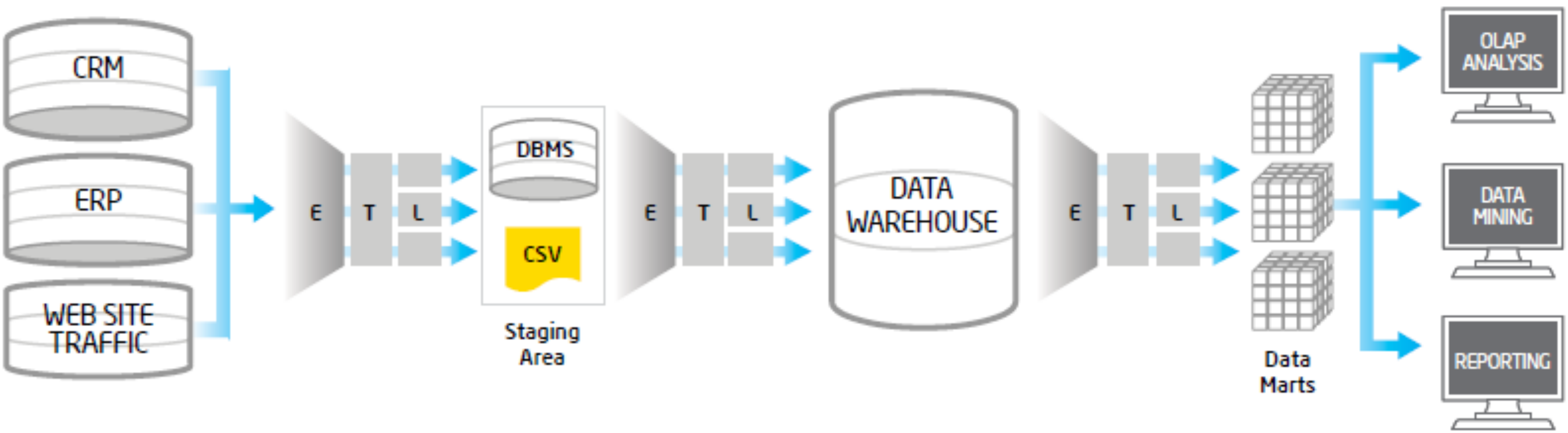
Source: http://hortonworks.com/hadoop/tez/

Big Data Solution



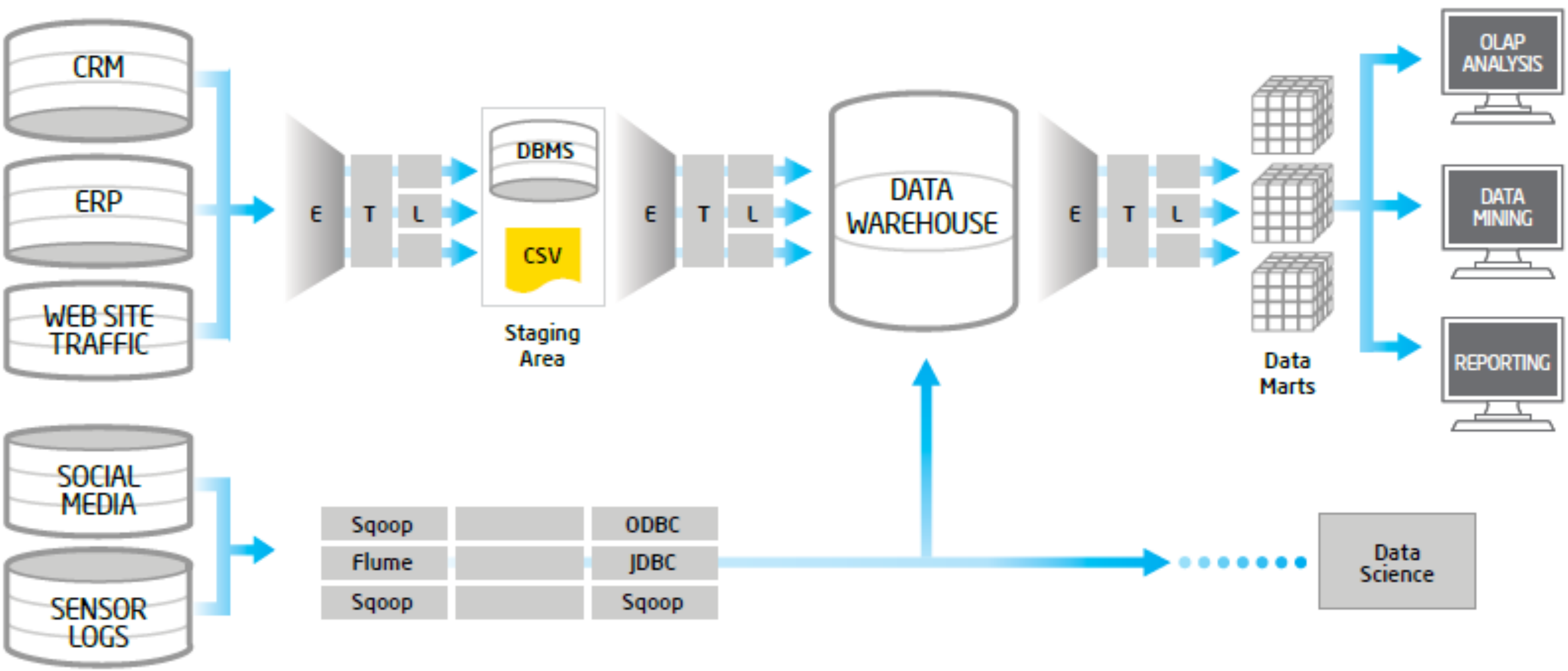
Source: http://www.newera-technologies.com/big-data-solution.html

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

Spark Ecosystem

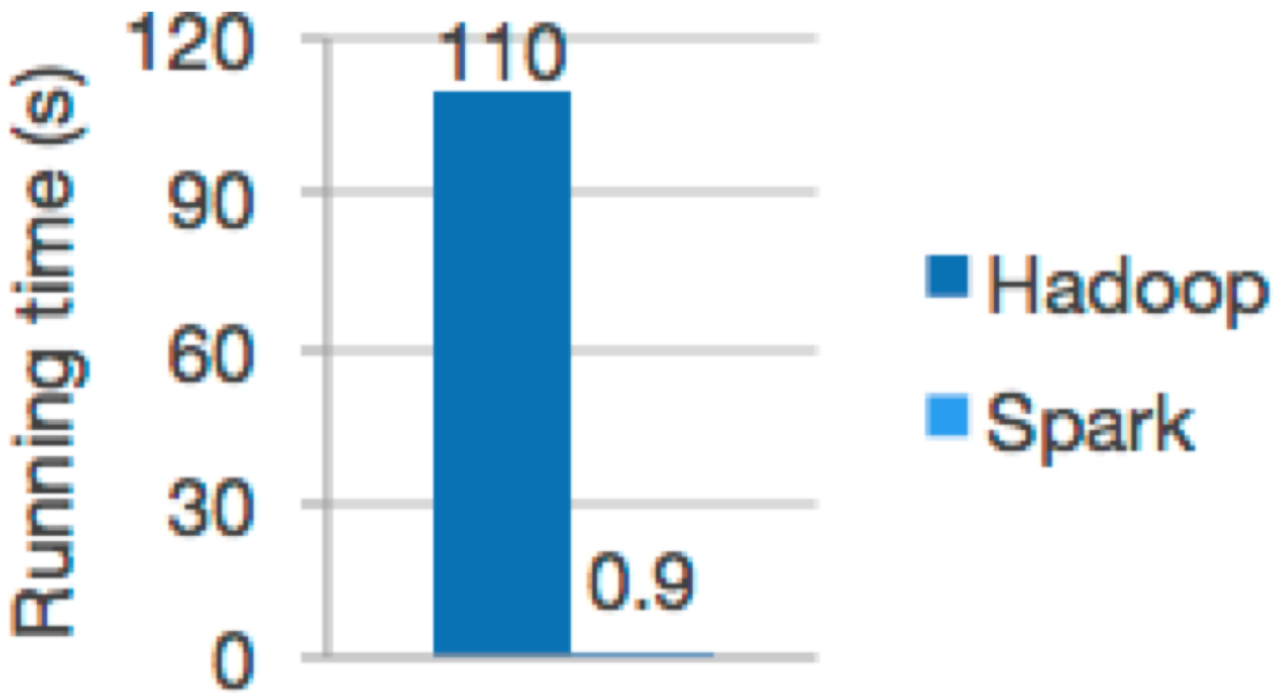
Lightning-fast cluster computing

Apache Spark

is a fast and general engine for large-scale data processing.

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

Logistic regression in Hadoop and Spark



Run programs up to 100x faster than Hadoop MapReduce in memory, or 10x faster on disk.

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

Ease of Use

- Write applications quickly in Java, Scala, Python, R.

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

Word count in Spark's Python API

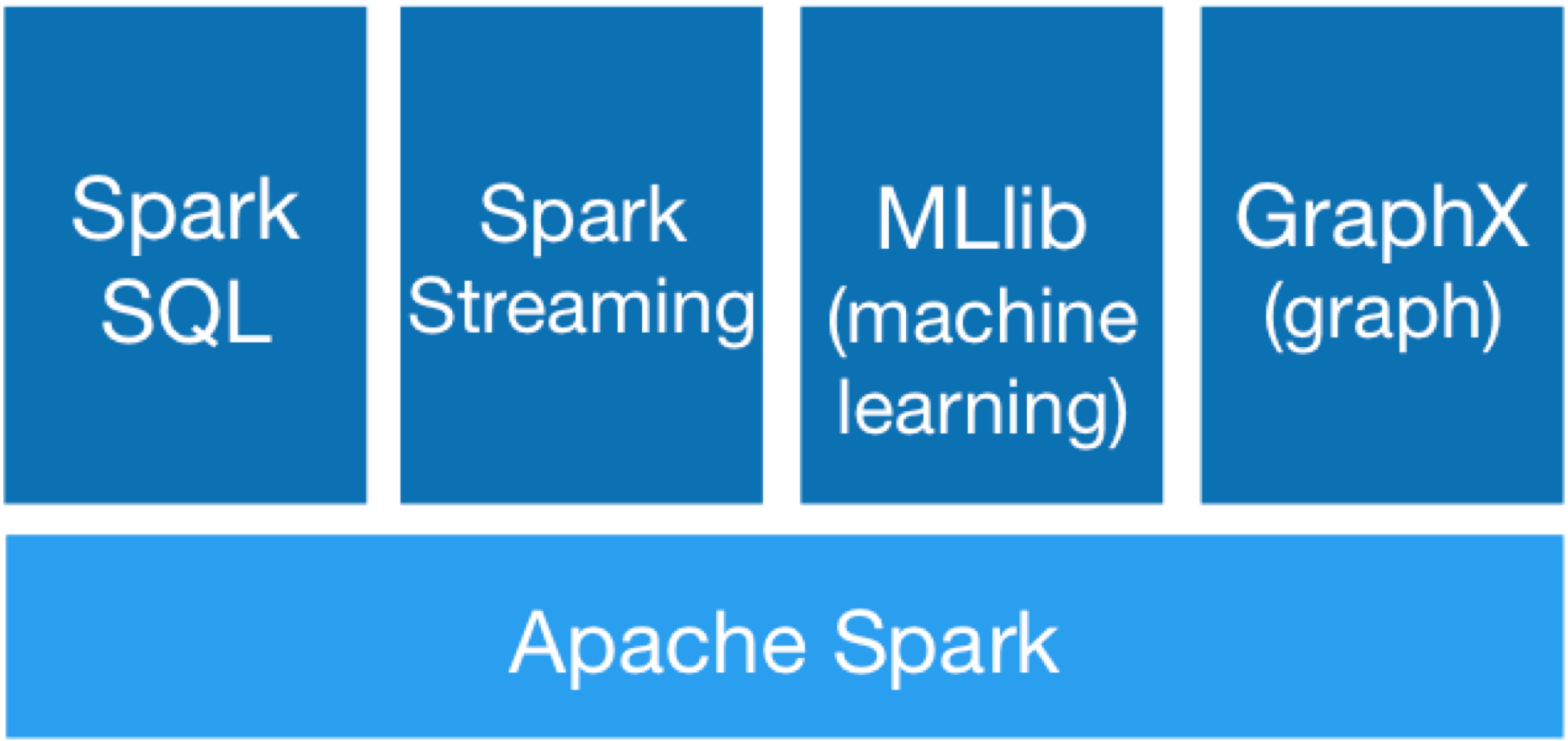
```
text_file = spark.textFile("hdfs://...")

text_file.flatMap(lambda line: line.split())
    .map(lambda word: (word, 1))
    .reduceByKey(lambda a, b: a+b)
```

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

Spark and Hadoop

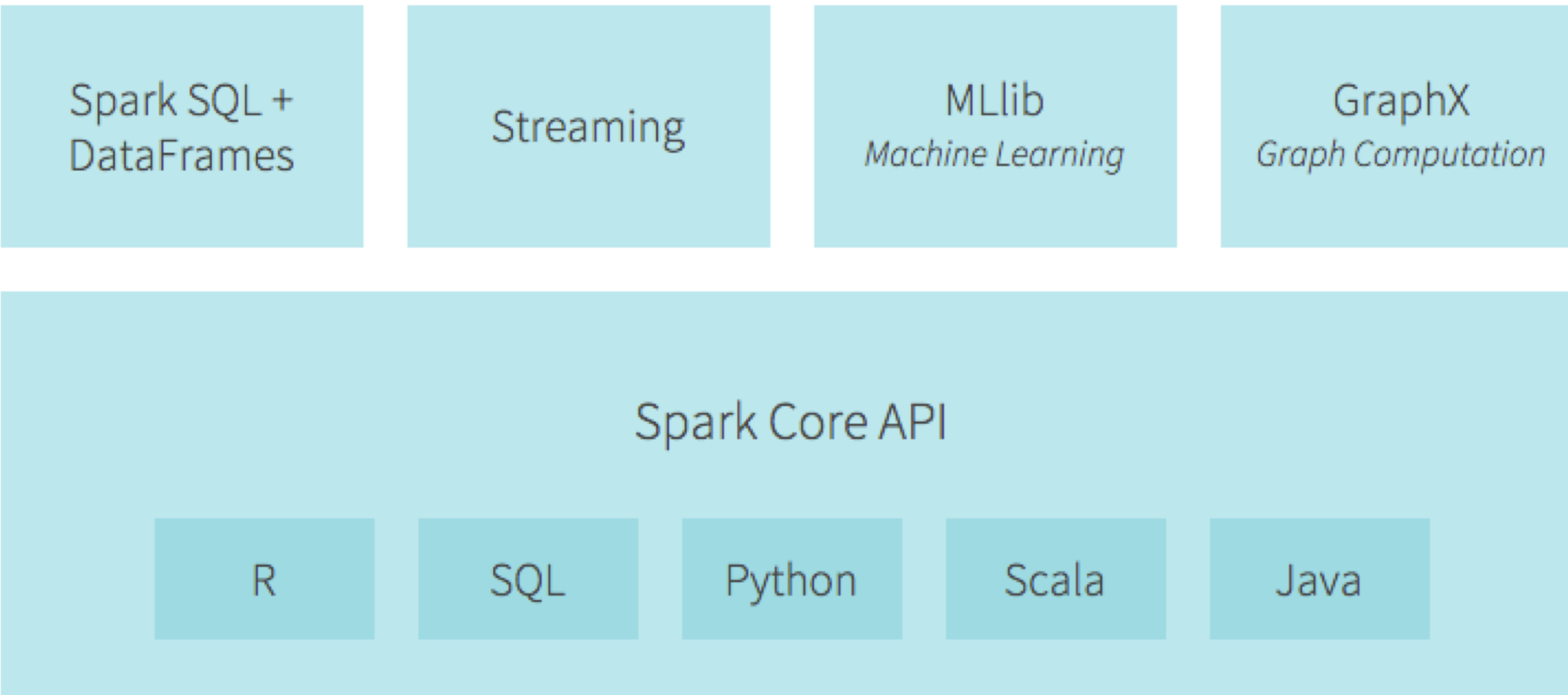
Source: http://spark.apache.org/

Spark Ecosystem



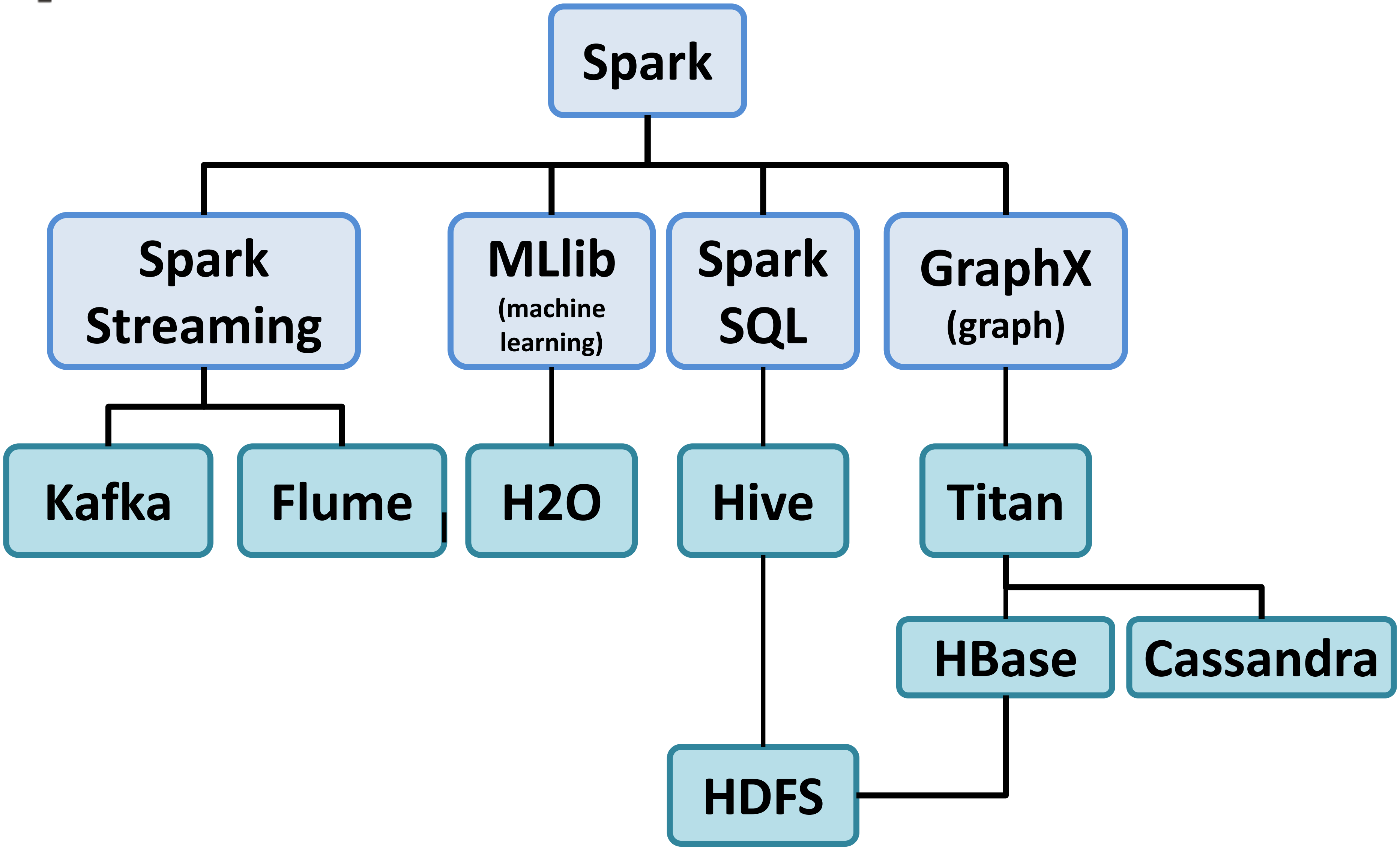
Source: http://spark.apache.org/

Spark Ecosystem



Source: https://databricks.com/spark/about

Spark Ecosystem



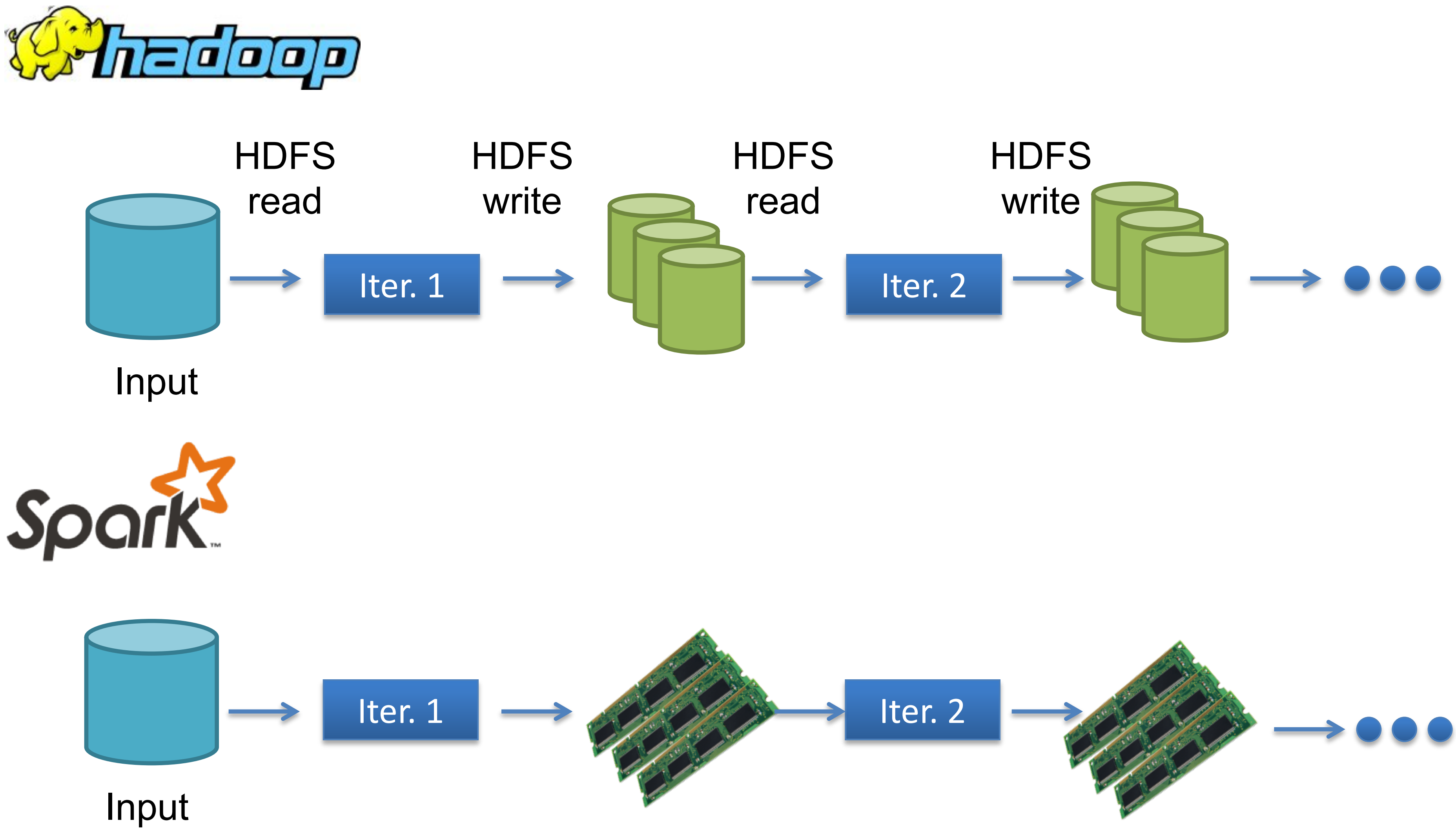
Source: Mike Frampton (2015), Mastering Apache Spark, Packt Publishing

SMACK Stack

- **Spark**
  - fast and general engine for distributed, large-scale data processing
- **Mesos**
  - cluster resource management system that provides efficient resource isolation and sharing across distributed applications
- **Akka**
  - a toolkit and runtime for building highly concurrent, distributed, and resilient message-driven applications on the JVM
- **Cassandra**
  - distributed, highly available database designed to handle large amounts of data across multiple datacenters
- **Kafka**
  - a high-throughput, low-latency distributed messaging system designed for handling real-time data feeds

Source: Anton Kirillov (2015), Data processing platforms architectures with Spark, Mesos, Akka, Cassandra and Kafka, Big Data AW Meetup

Hadoop vs. Spark



Source: Shiva Achari (2015), Hadoop Essentials - Tackling the Challenges of Big Data with Hadoop, Packt Publishing

Summary

- Big Data
- Artificial Intelligence
- Deep Learning
- Architectures of Big Data Analytics
- Data Mining Process
- Fundamental Big Data:
  MapReduce Paradigm,
  Hadoop and Spark Ecosystem

References

- EMC Education Services (2015), Data Science and Big Data Analytics: Discovering, Analyzing, Visualizing and Presenting Data, Wiley
- Stuart Russell and Peter Norvig (2016), Artificial Intelligence: A Modern Approach, 3rd Edition, Pearson International
- Peter C. Verhoef and Edwin Kooge (2016), Creating Value with Big Data Analytics: Making Smarter Marketing Decisions, Routledge
- Shiva Achari (2015), Hadoop Essentials - Tackling the Challenges of Big Data with Hadoop, Packt Publishing
- Mike Frampton (2015), Mastering Apache Spark, Packt Publishing
- Deepak Ramanathan (2014), SAS Modernization architectures - Big Data Analytics, http://www.slideshare.net/deepakramanathan/sas-modernization-architectures-big-data-analytics