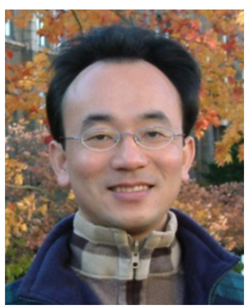
Tamkang University
淡江大學

Tamkang University

Social Computing and Big Data Analytics
社群運算與大數據分析

Data Science and Big Data Analytics:
Discovering, Analyzing, Visualizing and Presenting Data
(資料科學與大數據分析：
探索、分析、視覺化與呈現資料)

1052SCBDA02
MIS MBA (M2226) (8606)
Wed, 8,9, (15:10-17:00) (B505)



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/

2017-02-22

課程大綱 (Syllabus)

週次 (Week)　日期 (Date)　內容 (Subject/Topics)

1　2017/02/15　Course Orientation for Social Computing and Big Data Analytics
(社群運算與大數據分析課程介紹)

2　2017/02/22　Data Science and Big Data Analytics: Discovering, Analyzing, Visualizing and Presenting Data
(資料科學與大數據分析：探索、分析、視覺化與呈現資料)

3　2017/03/01　Fundamental Big Data: MapReduce Paradigm, Hadoop and Spark Ecosystem
(大數據基礎：MapReduce典範、Hadoop與Spark生態系統)

課程大綱 (Syllabus)

週次 (Week)	日期 (Date)	內容 (Subject/Topics)
4	2017/03/08	Big Data Processing Platforms with SMACK: Spark, Mesos, Akka, Cassandra and Kafka (大數據處理平台SMACK： Spark, Mesos, Akka, Cassandra, Kafka)
5	2017/03/15	Big Data Analytics with Numpy in Python (Python Numpy 大數據分析)
6	2017/03/22	Finance Big Data Analytics with Pandas in Python (Python Pandas 財務大數據分析)
7	2017/03/29	Text Mining Techniques and Natural Language Processing (文字探勘分析技術與自然語言處理)
8	2017/04/05	Off-campus study (教學行政觀摩日)

課程大綱 (Syllabus)

週次 (Week)　日期 (Date)　內容 (Subject/Topics)

9　2017/04/12　Social Media Marketing Analytics
(社群媒體行銷分析)

10　2017/04/19　期中報告 (Midterm Project Report)

11　2017/04/26　Deep Learning with Theano and Keras in Python
(Python Theano 和 Keras 深度學習)

12　2017/05/03　Deep Learning with Google TensorFlow
(Google TensorFlow 深度學習)

13　2017/05/10　Sentiment Analysis on Social Media with Deep Learning
(深度學習社群媒體情感分析)

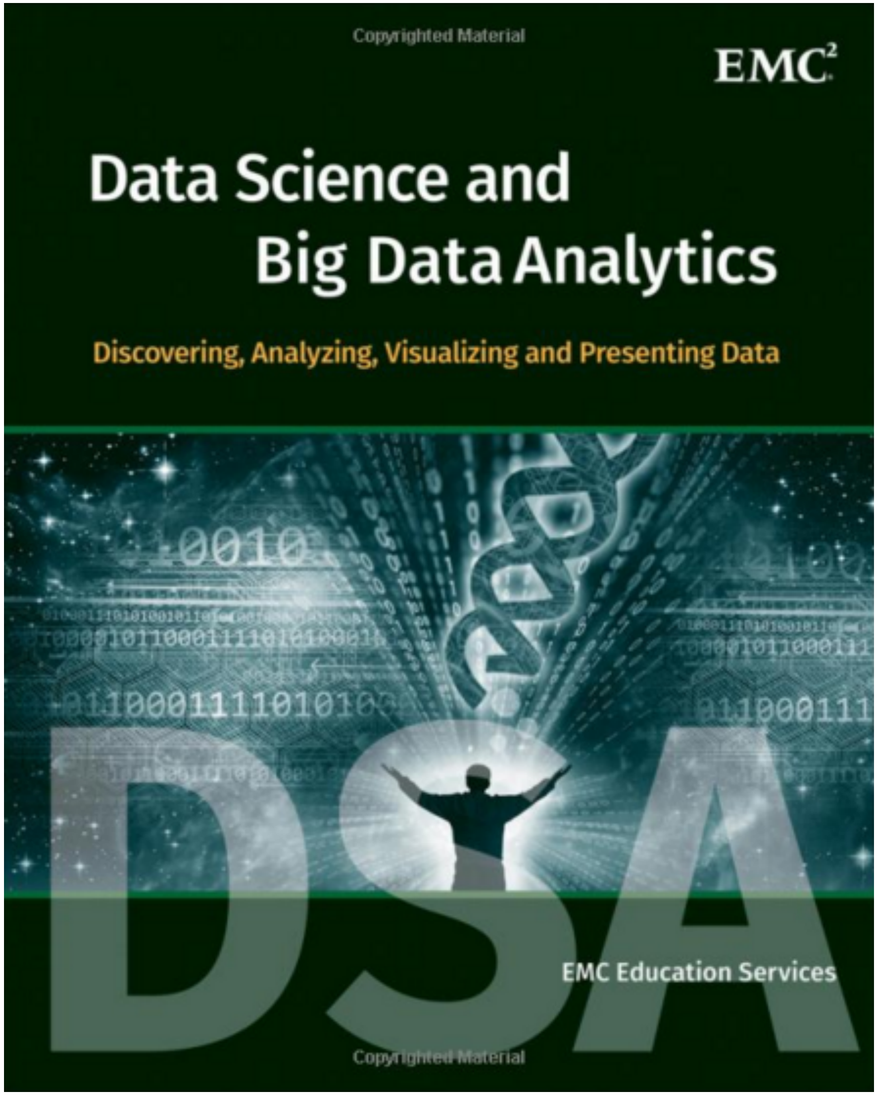
課程大綱 (Syllabus)

週次 (Week)	日期 (Date)	內容 (Subject/Topics)
14	2017/05/17	Social Network Analysis (社會網絡分析)
15	2017/05/24	Measurements of Social Network (社會網絡量測)
16	2017/05/31	Tools of Social Network Analysis (社會網絡分析工具)
17	2017/06/07	Final Project Presentation I (期末報告 I)
18	2017/06/14	Final Project Presentation II (期末報告 II)

2017/02/22

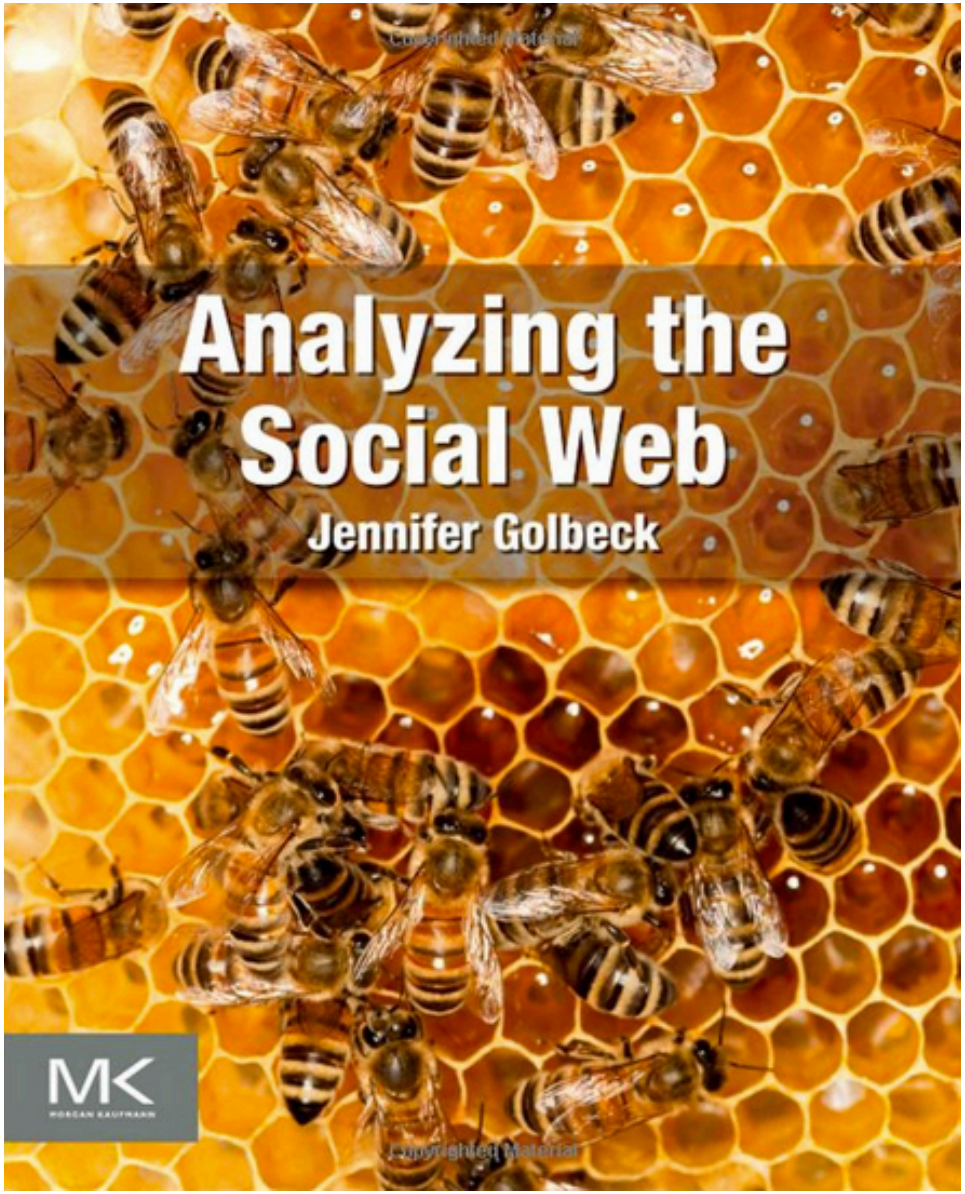
Data Science and Big Data Analytics: Discovering, Analyzing, Visualizing and Presenting Data (資料科學與大數據分析：探索、分析、視覺化與呈現資料)

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



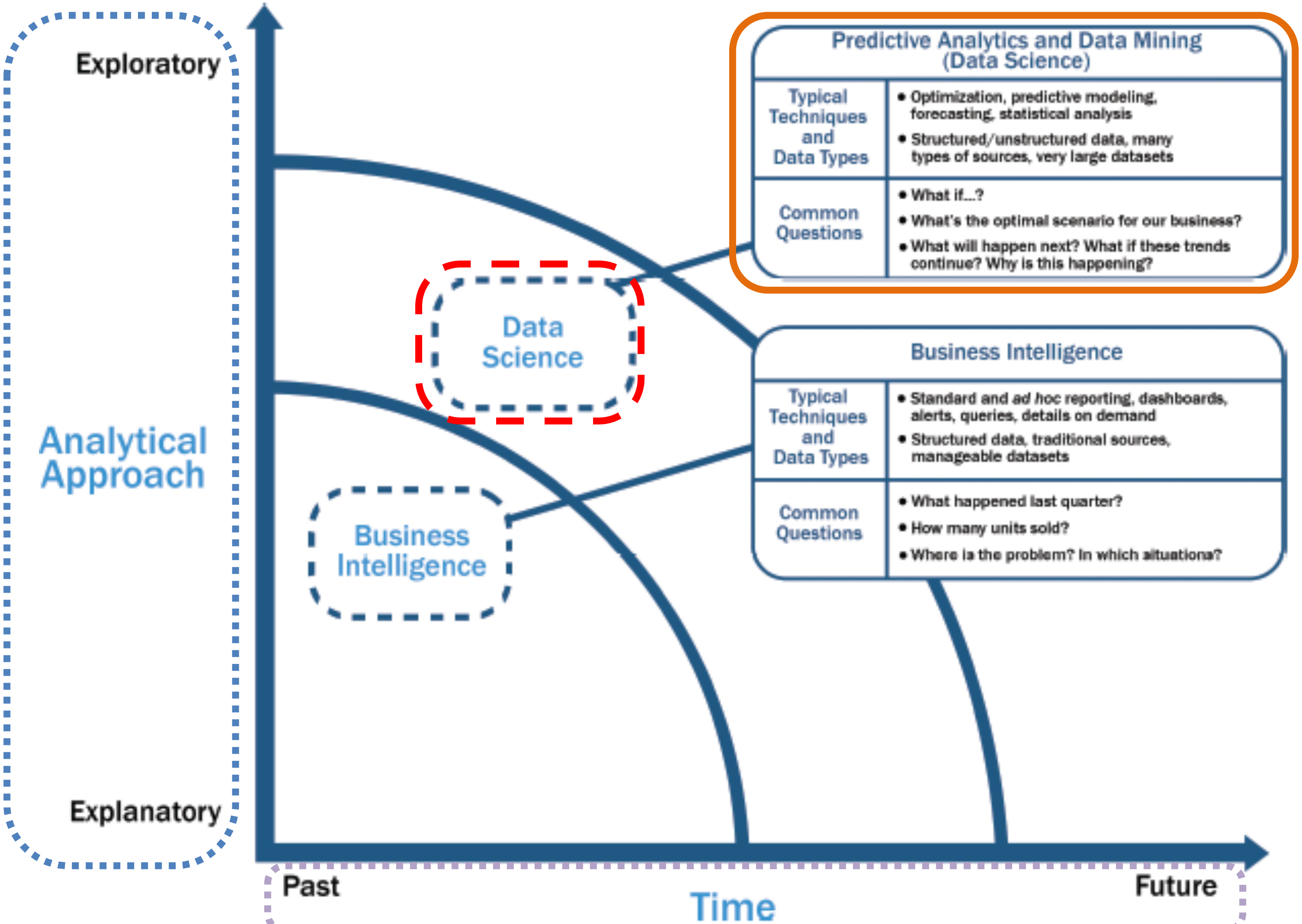
Source: http://www.amazon.com/Data-Science-Big-Analytics-Discovering/dp/111887613X

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



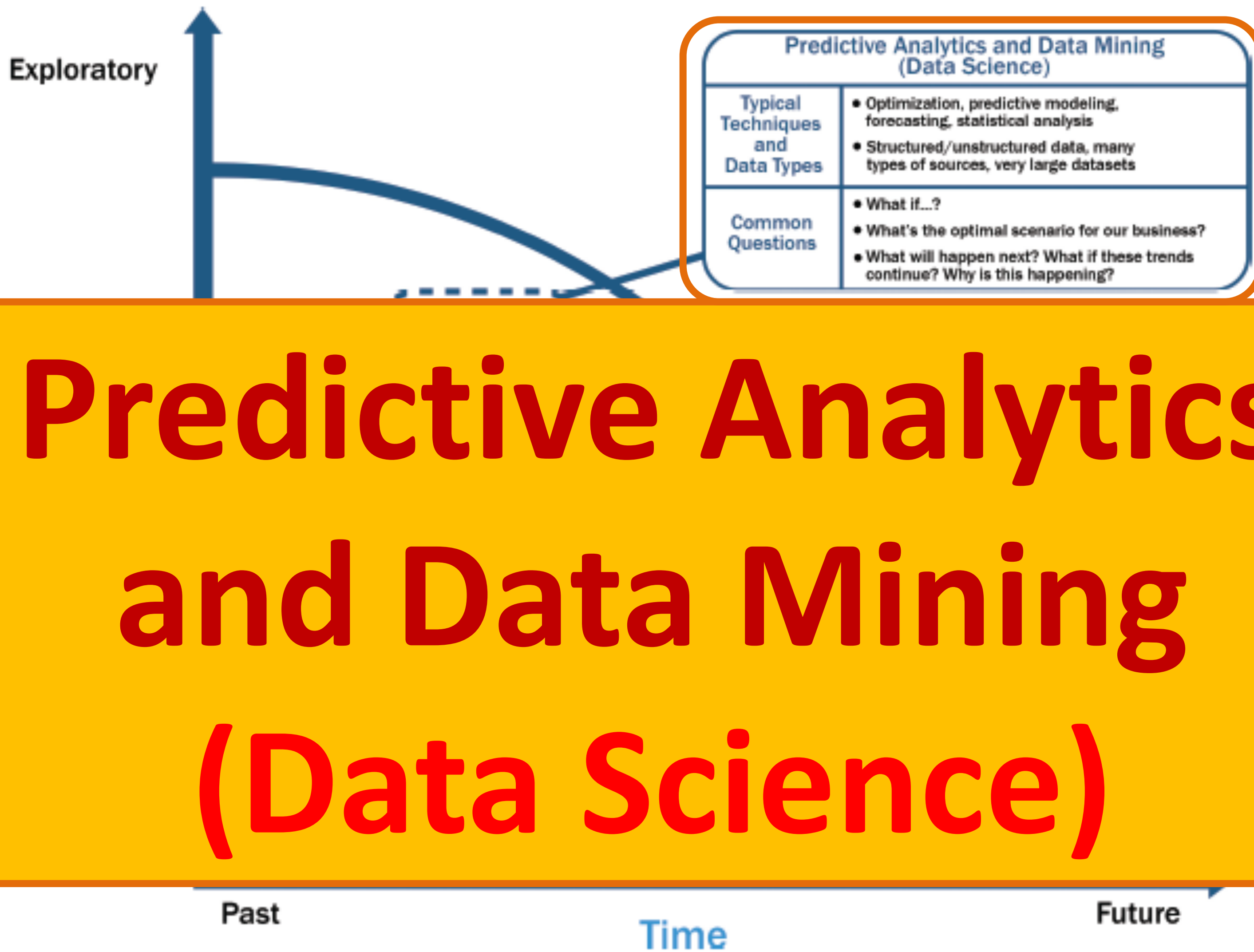
Source: http://www.amazon.com/Analyzing-Social-Web-Jennifer-Golbeck/dp/0124055311

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)

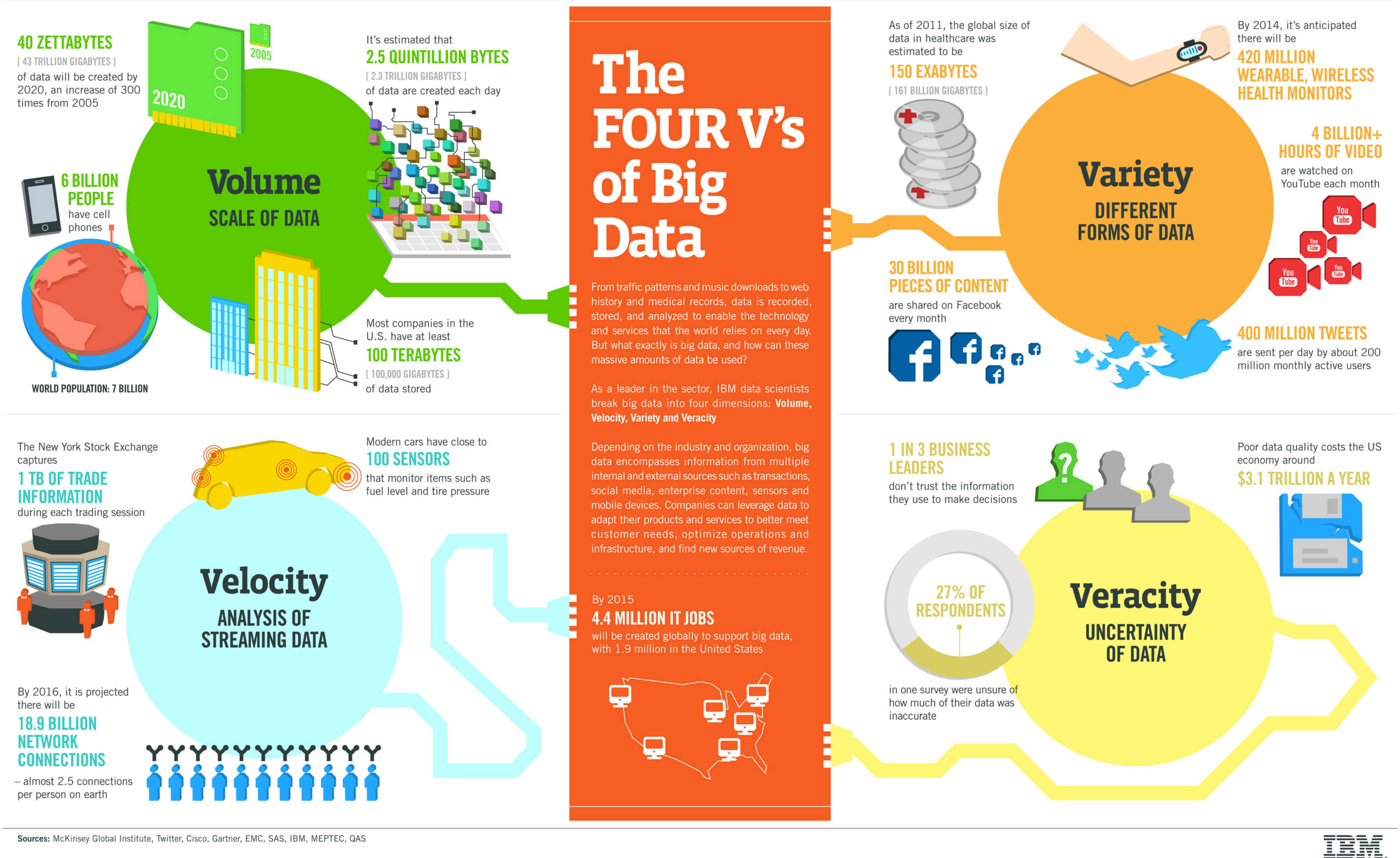
Structured/unstructured data, many types of sources, very large datasets

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

Big Data 4 V



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

Value

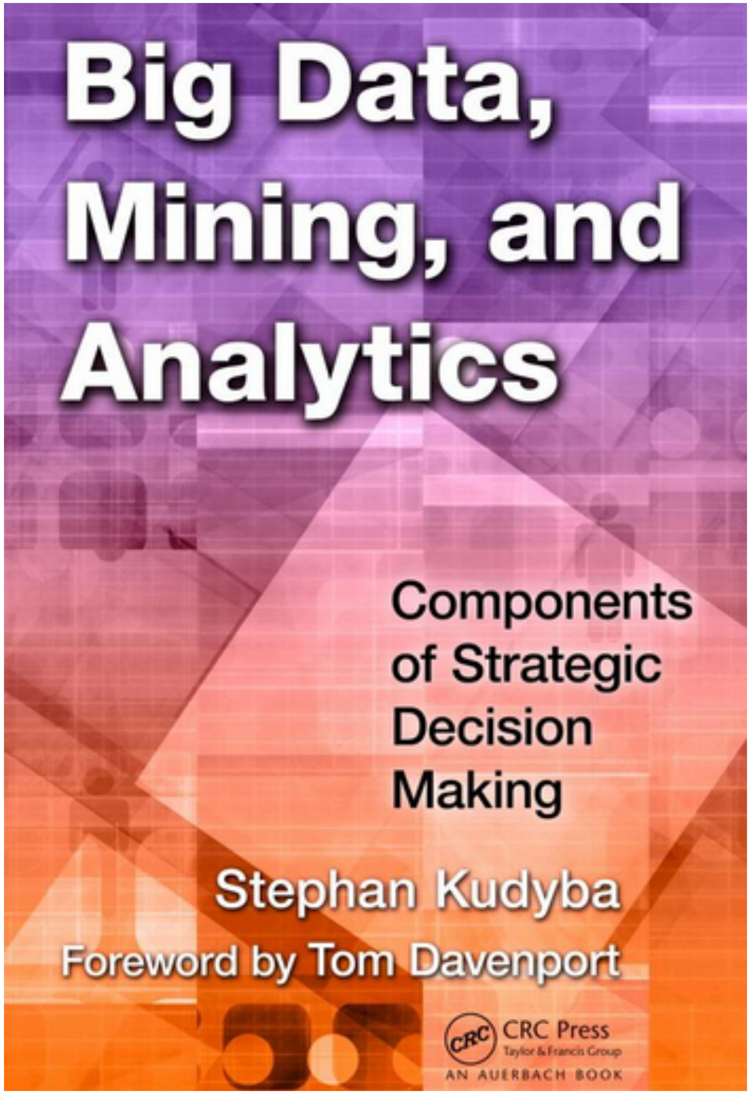
Big Data Analytics

and

Data Mining

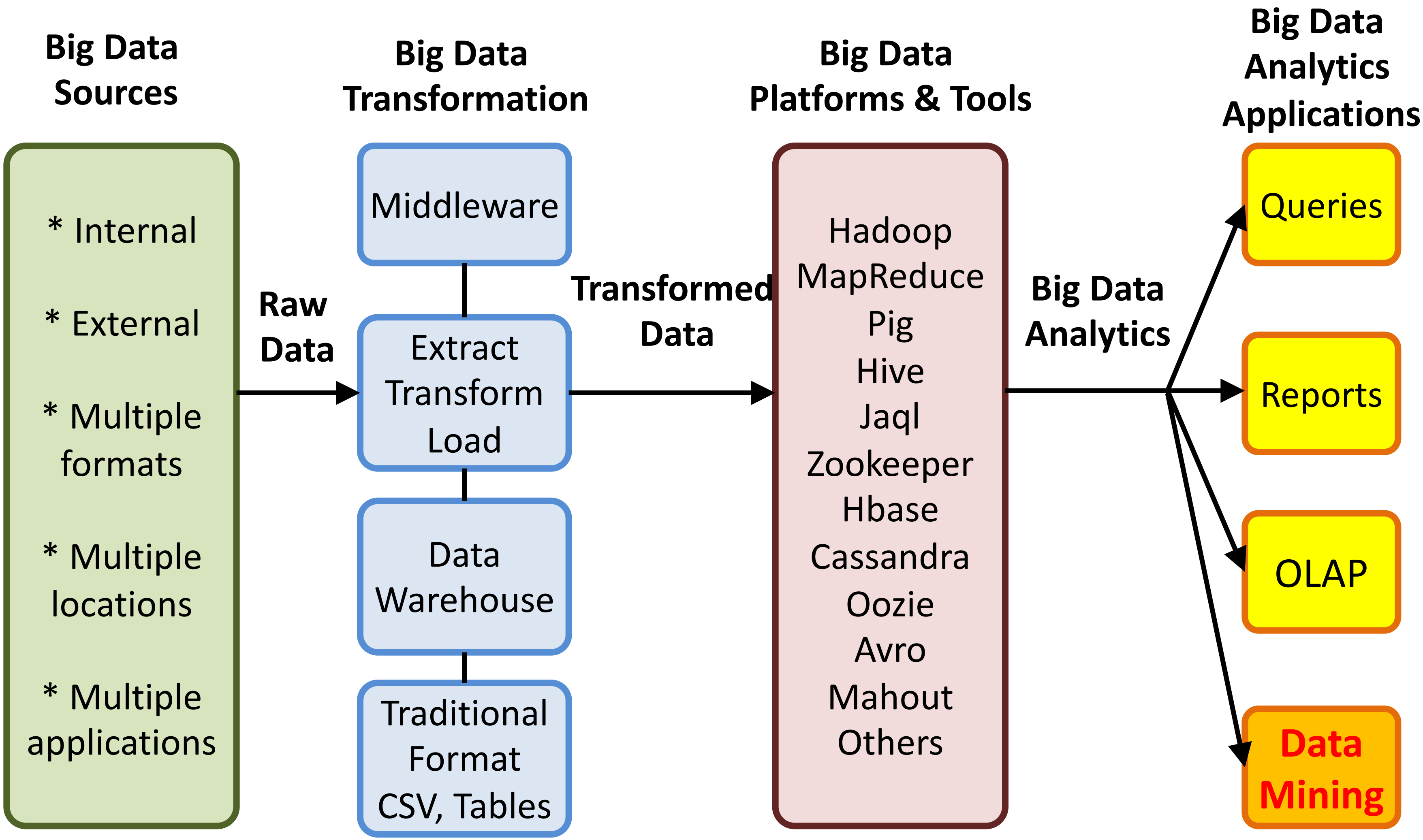
Stephan Kudyba (2014),

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



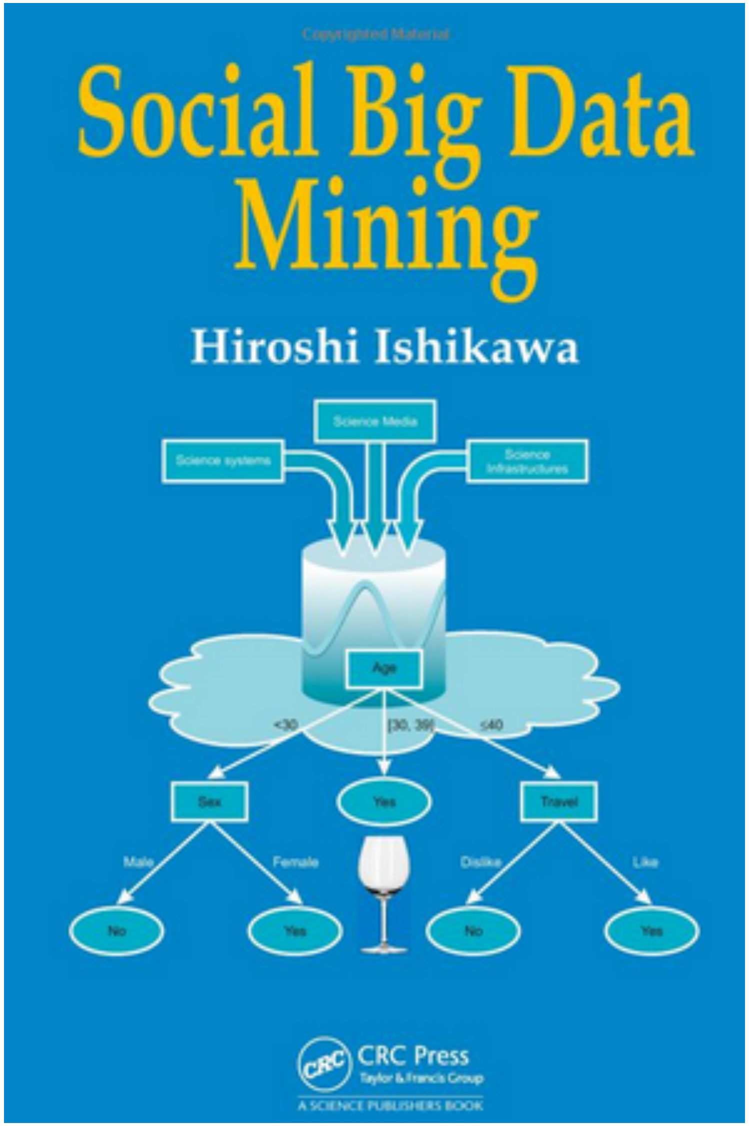
Source: Stephan Kudyba (2014), Big Data, Mining, and Analytics: Components of Strategic Decision Making, Auerbach Publications

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)



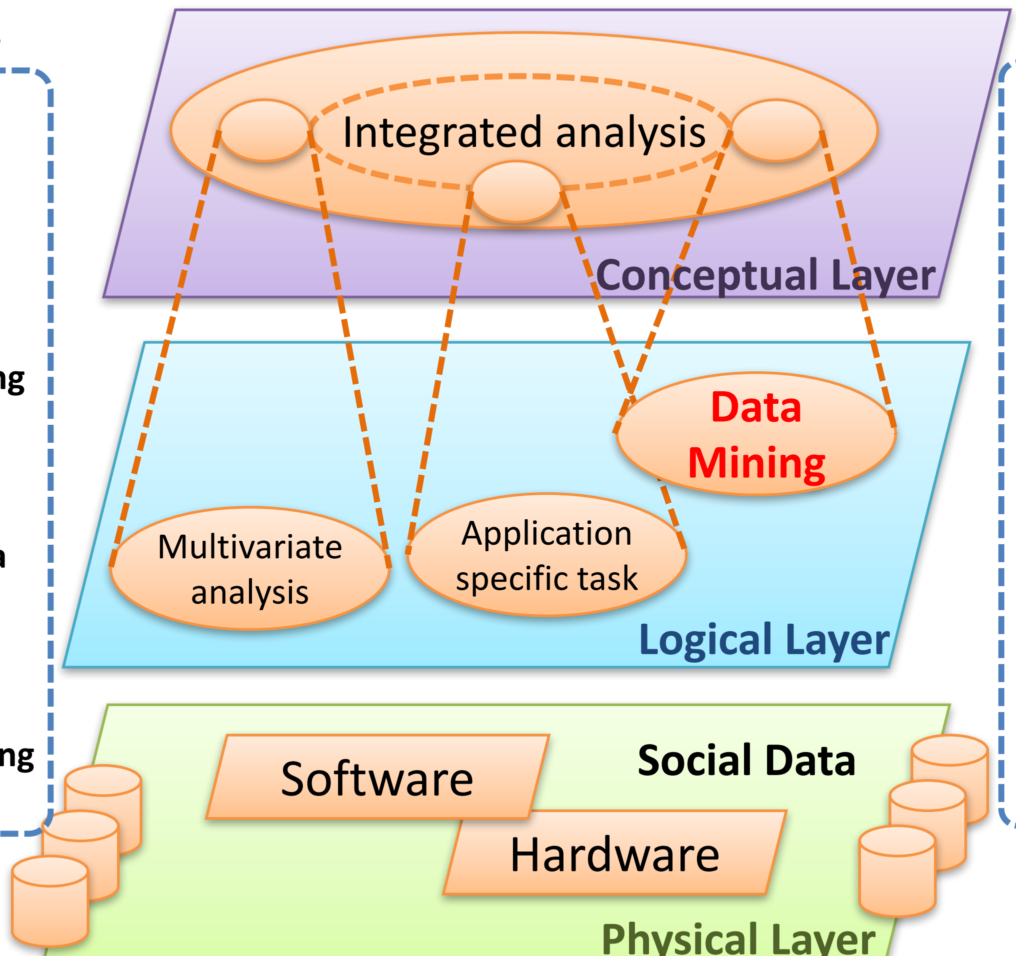
Source: http://www.amazon.com/Social-Data-Mining-Hiroshi-Ishikawa/dp/149871093X

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

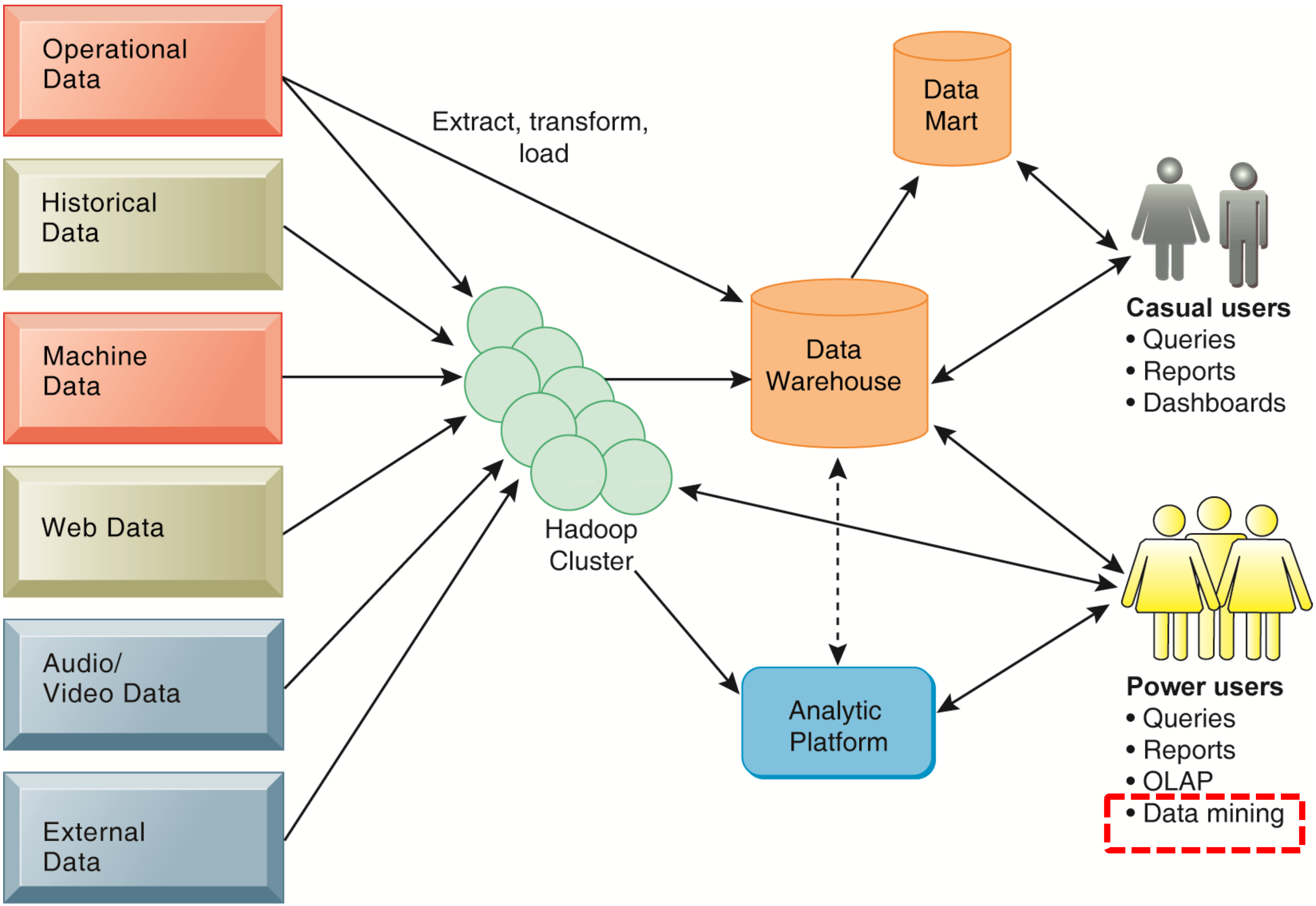


Analysts

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

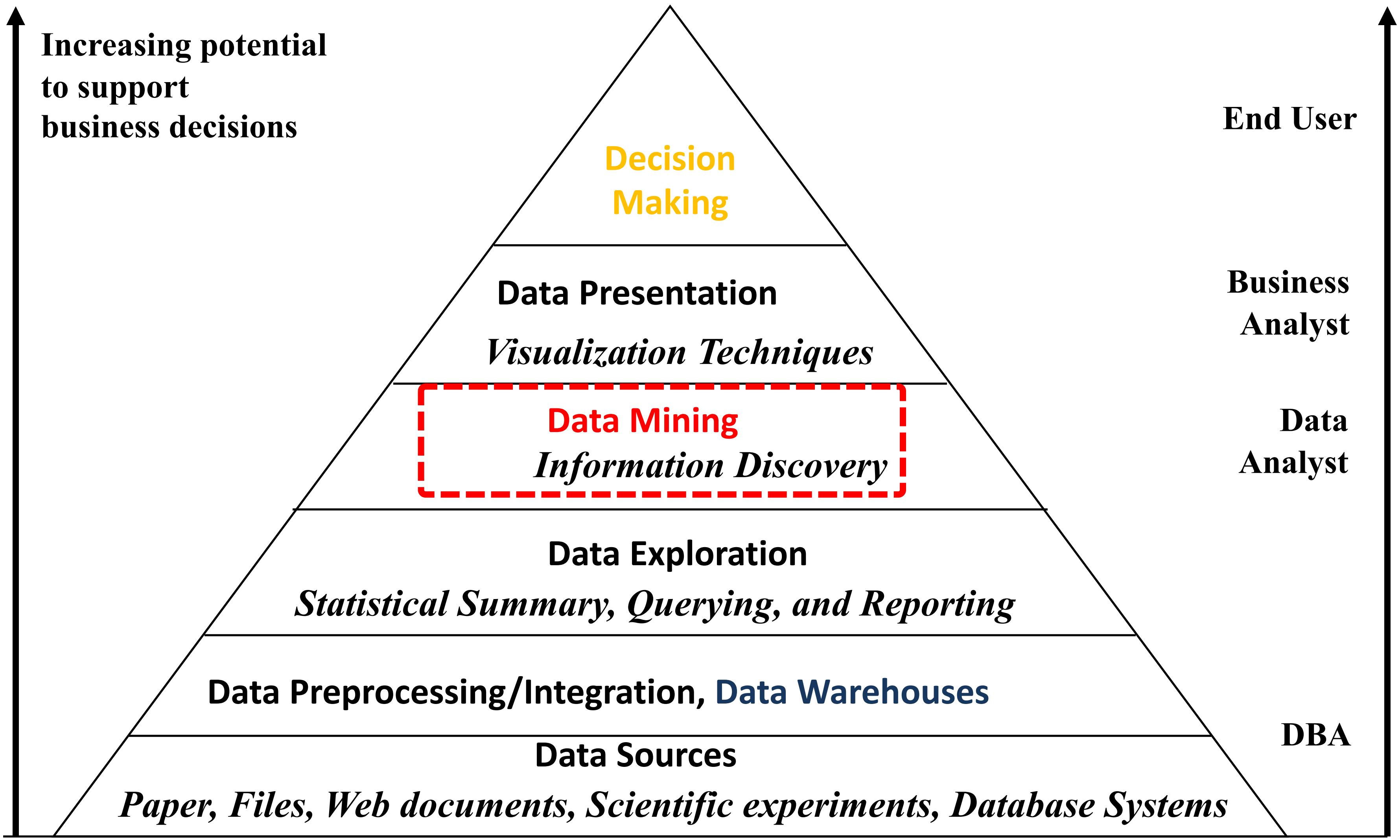
Source: Hiroshi Ishikawa (2015), Social Big Data Mining, CRC Press

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

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)

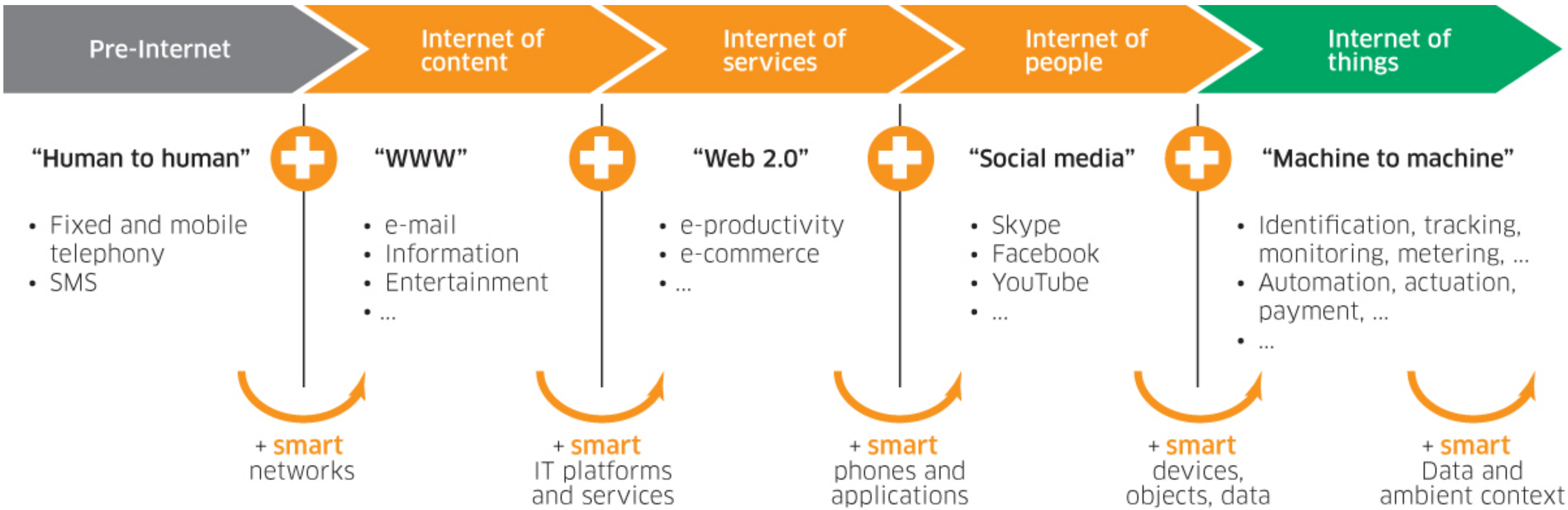
Source: Jiawei Han and Micheline Kamber (2011), Data Mining: Concepts and Techniques, Third Edition, Elsevier

Big Data Analysis

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

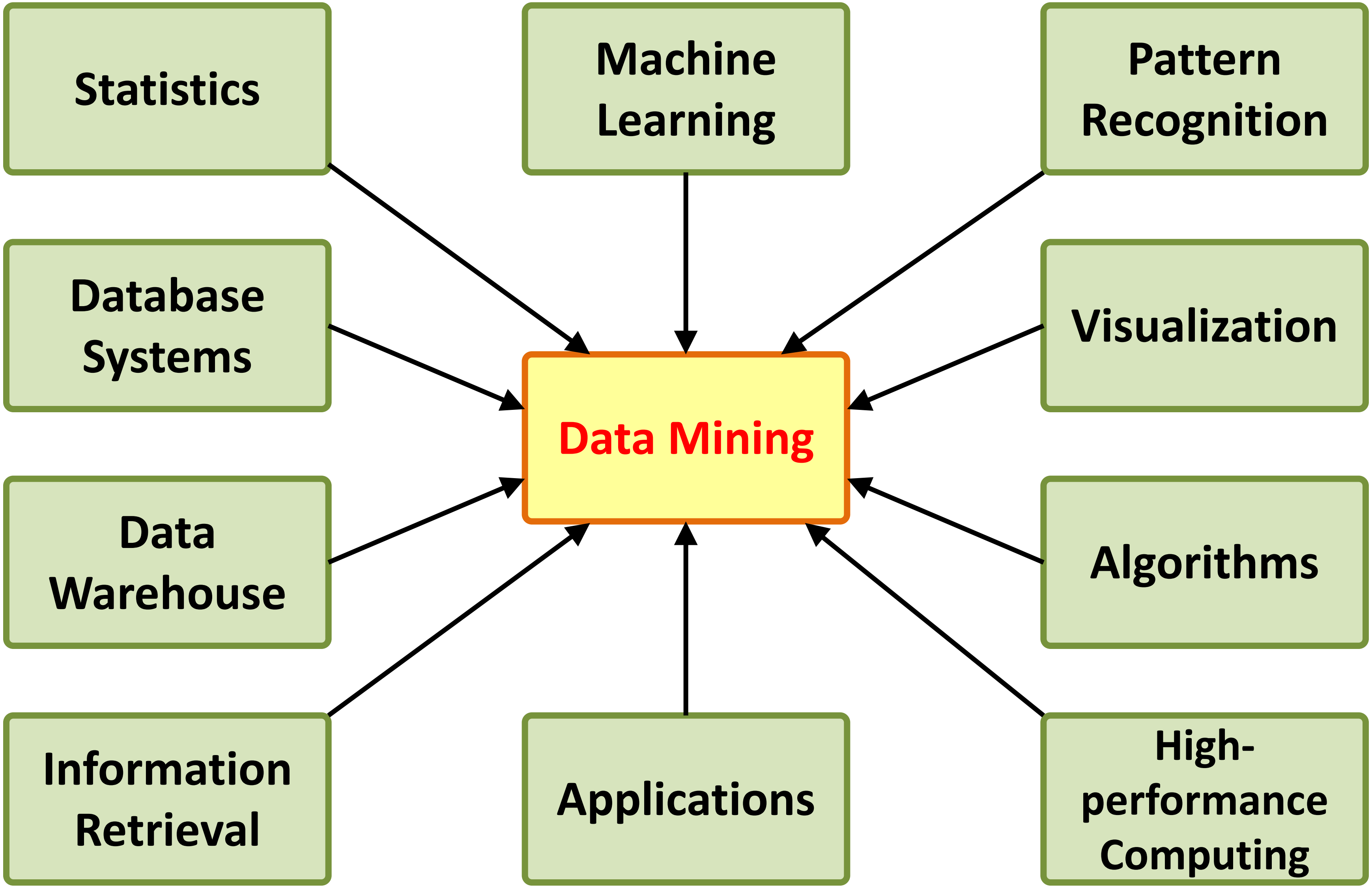
Internet Evolution

Internet of People (IoP): Social Media
Internet of Things (IoT): Machine to Machine

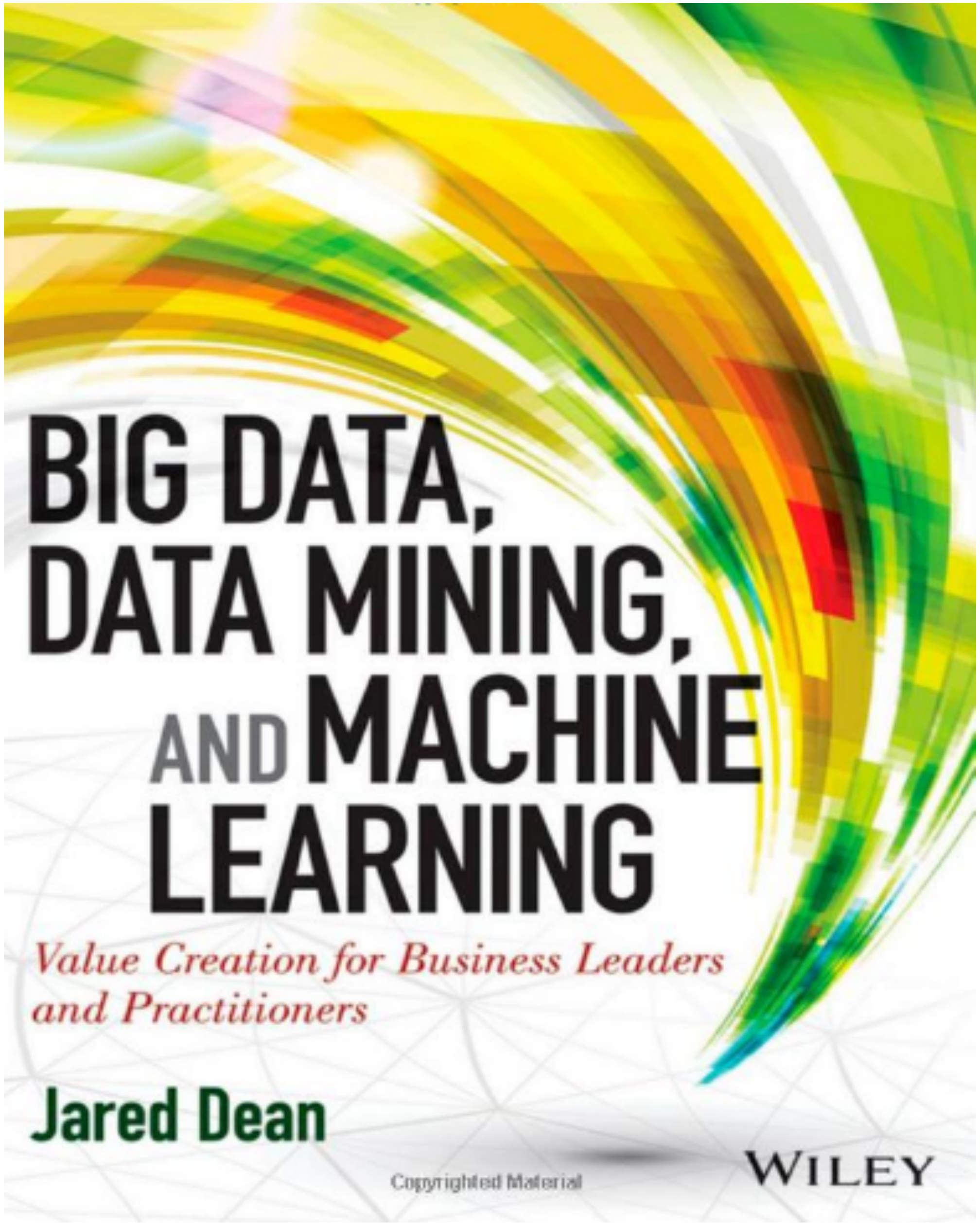


Source: Marc Jadoul (2015), The IoT: The next step in internet evolution, March 11, 2015
http://www2.alcatel-lucent.com/techzine/iot-internet-of-things-next-step-evolution/

Data Mining Technologies



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

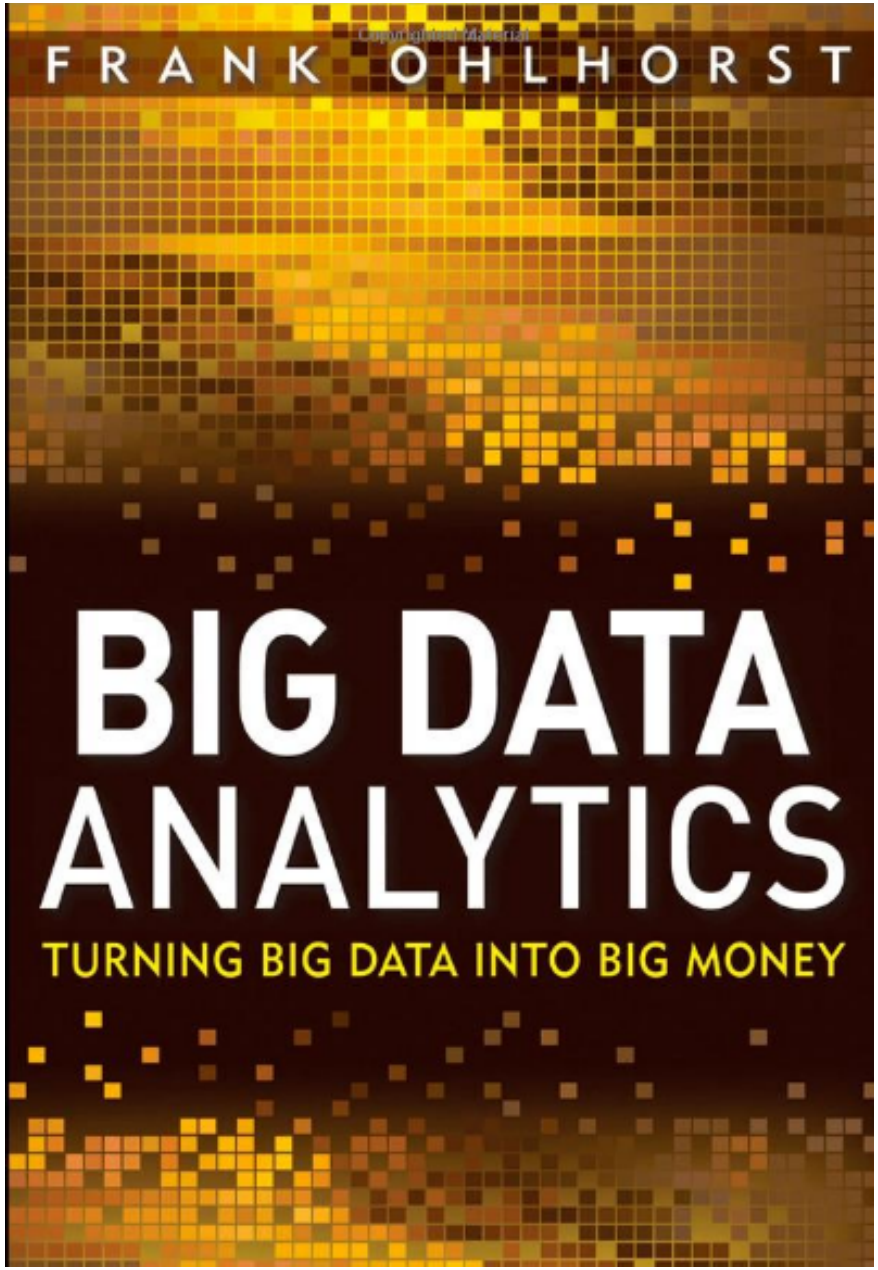


Source: http://www.amazon.com/Data-Mining-Machine-Learning-Practitioners/dp/1118618041

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

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

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?"	
Predictive Modeling / Forecasting	"What will happen next?"	**Predictive Analytics**
Statistical Modeling	"Why is this happening?"	
Alerts	"What actions are needed?"	**Descriptive Analytics**
Query / Drill Down	"What exactly is the problem?"	
Ad hoc Reports / Scorecards	"How many, how often, where?"	
Standard Report	"What happened?"	

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

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

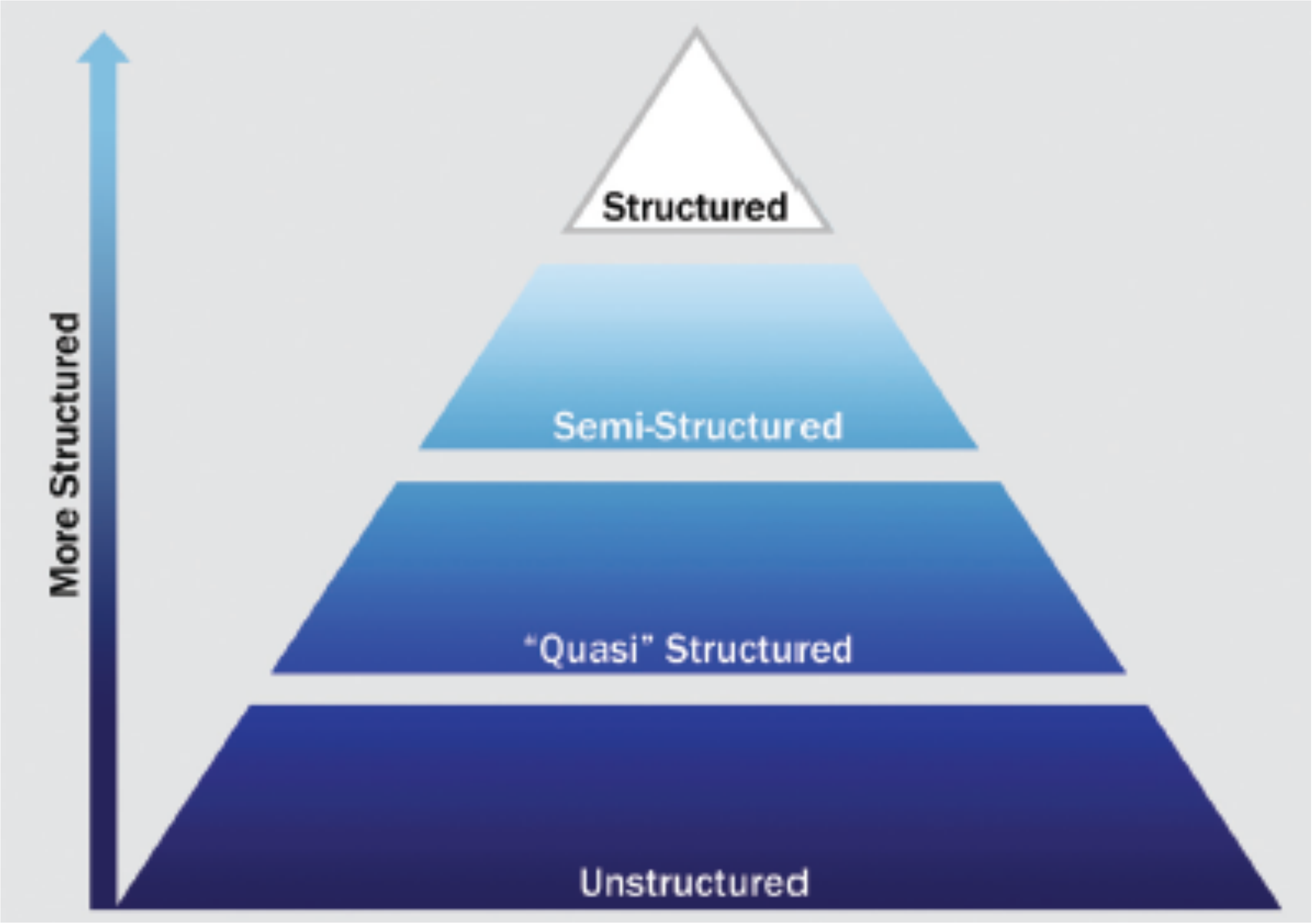
When Jonathan Goldman arrived for work in June 2006 at LinkedIn, the business networking site, the place still felt like a start-up. The company had just under 8 million accounts, and the number was growing quickly as existing members invited their friends and colleagues to join. But users weren't seeking 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*

Big Data

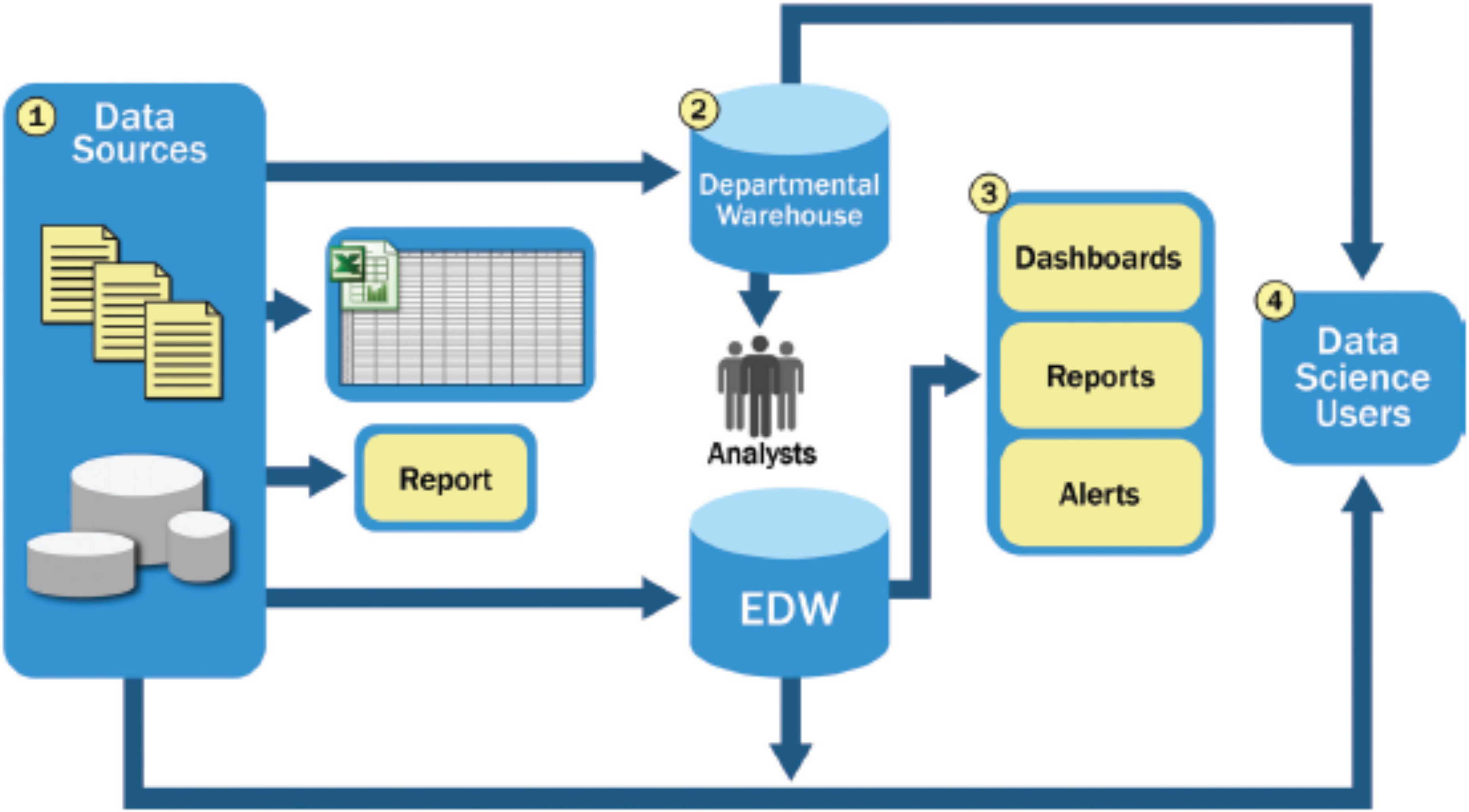
Source: EMC Education Services, Data Science and Big Data Analytics: Discovering, Analyzing, Visualizing and Presenting Data, Wiley, 2015

Big Data Growth is increasingly unstructured



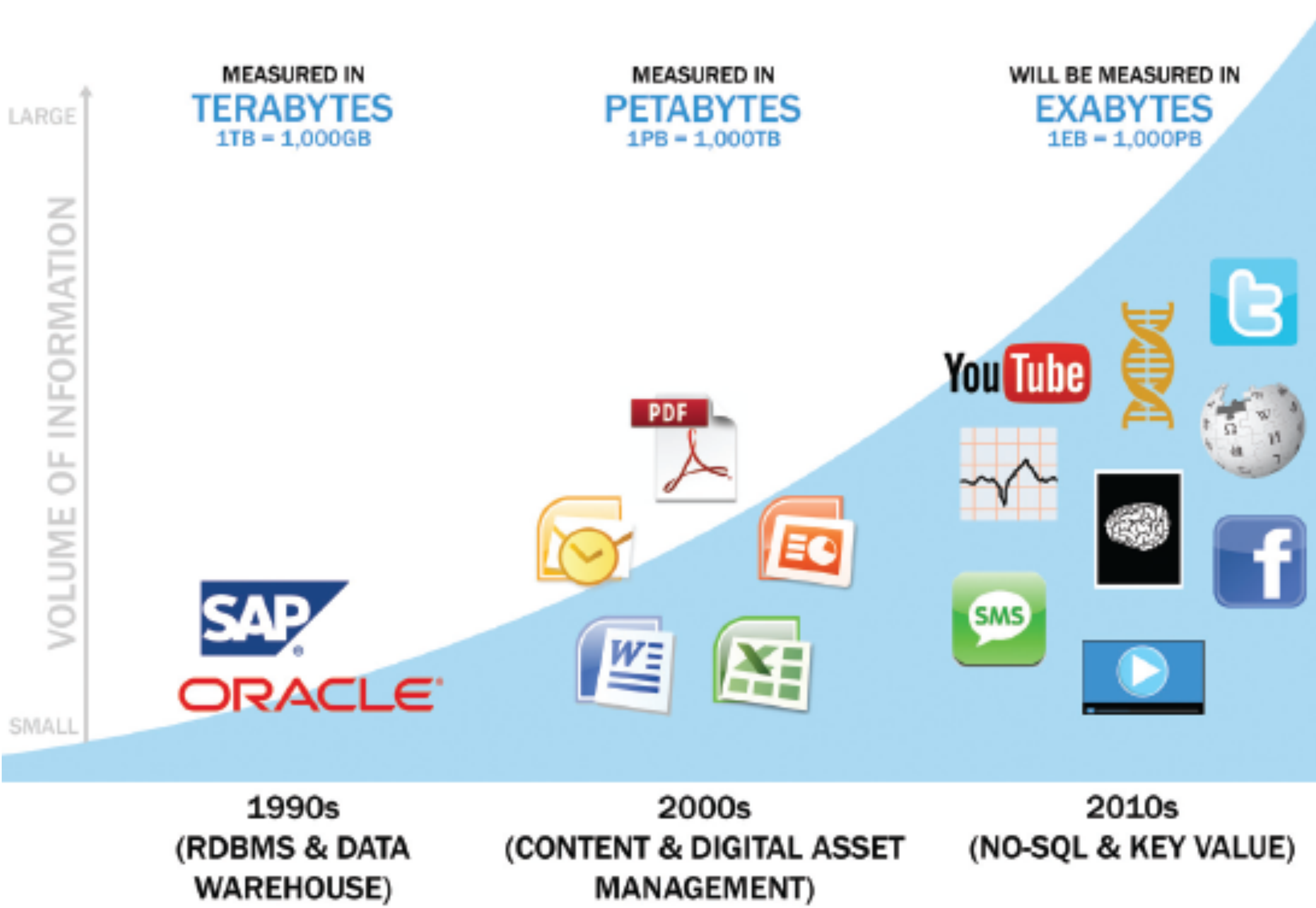
Source: EMC Education Services, Data Science and Big Data Analytics: Discovering, Analyzing, Visualizing and Presenting Data, Wiley, 2015

Typical Analytic Architecture



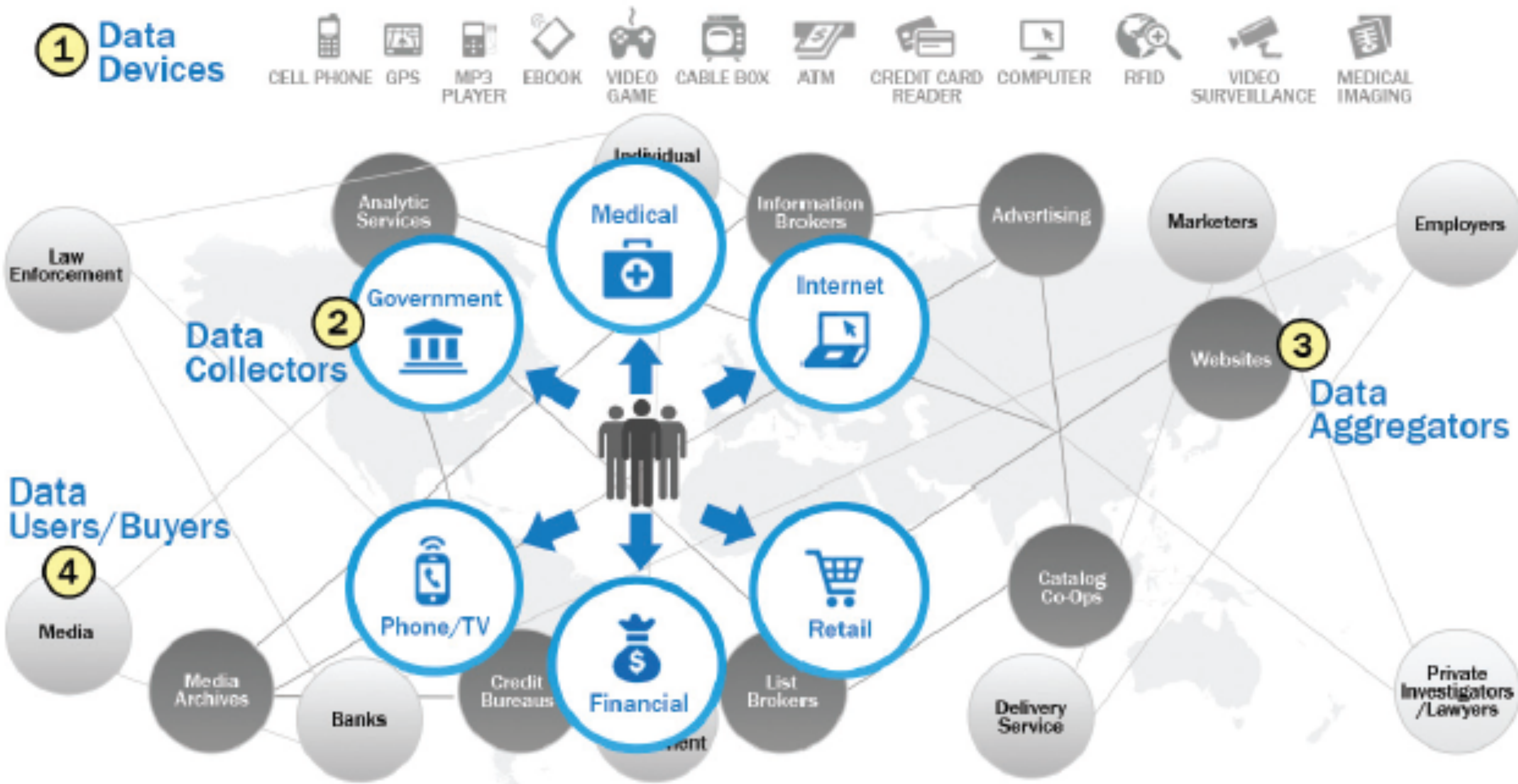
Source: EMC Education Services, Data Science and Big Data Analytics: Discovering, Analyzing, Visualizing and Presenting Data, Wiley, 2015

Data Evolution and the Rise of Big Data Sources



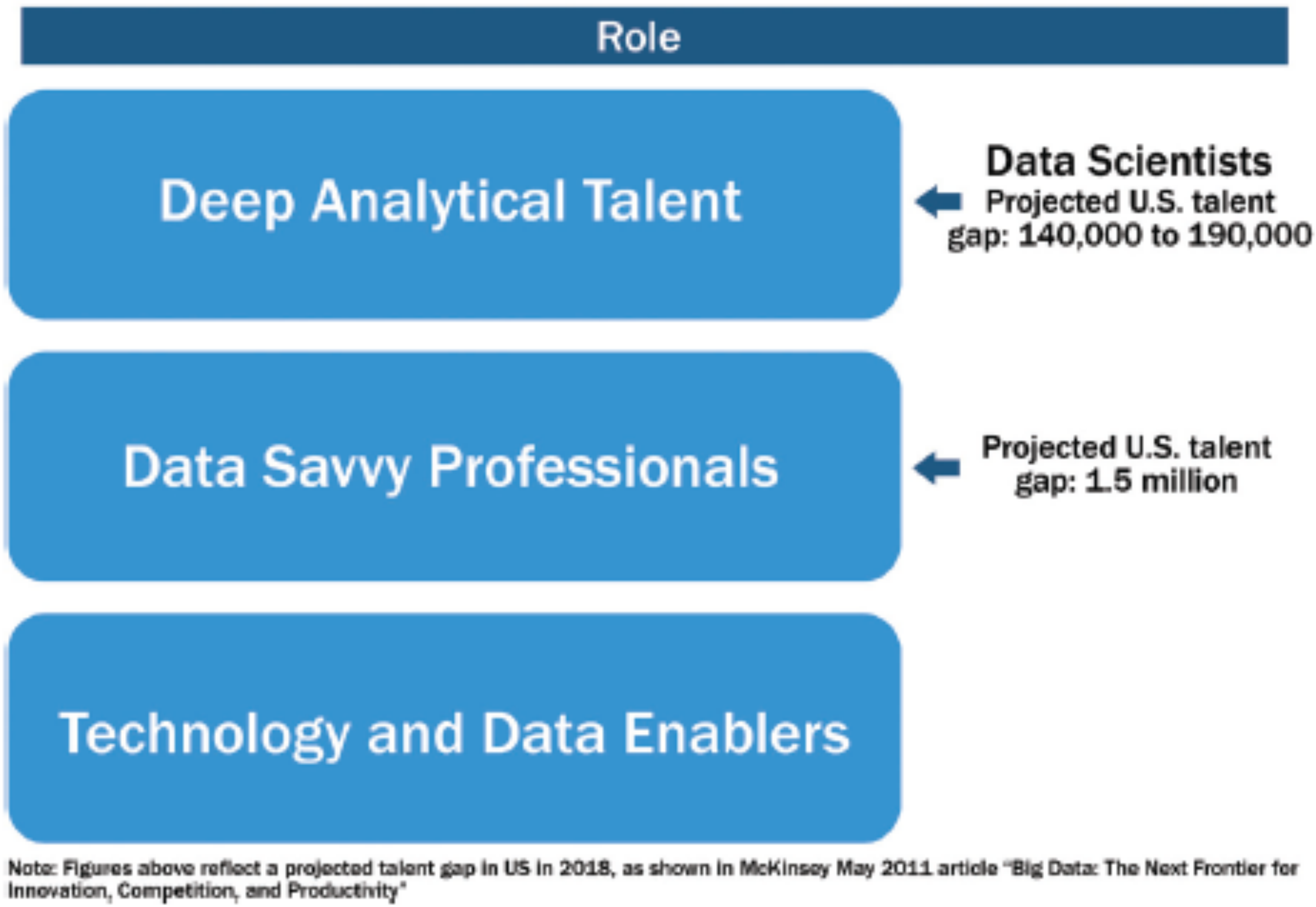
Source: EMC Education Services, Data Science and Big Data Analytics: Discovering, Analyzing, Visualizing and Presenting Data, Wiley, 2015

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



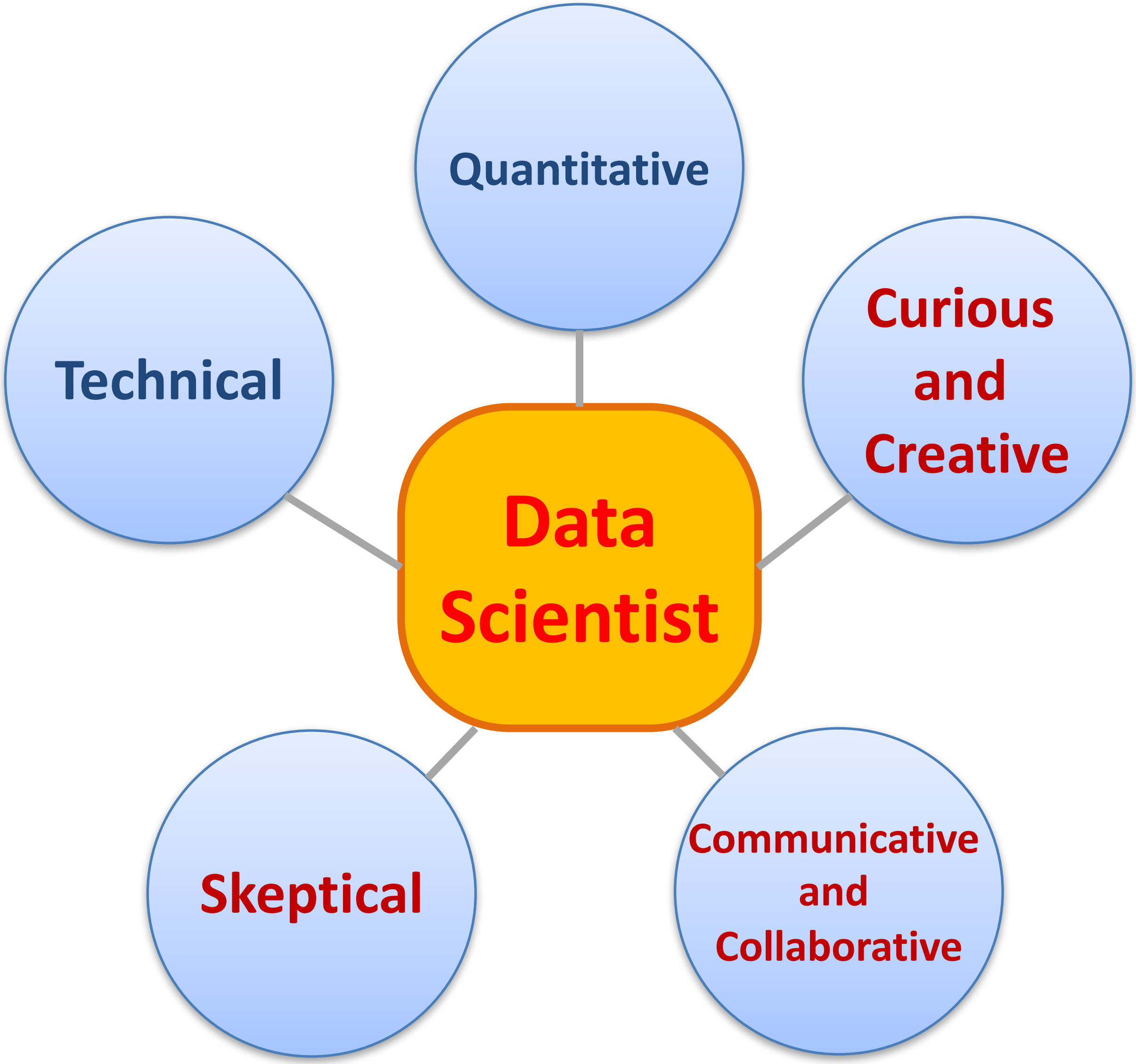
Source: EMC Education Services, Data Science and Big Data Analytics: Discovering, Analyzing, Visualizing and Presenting Data, Wiley, 2015

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

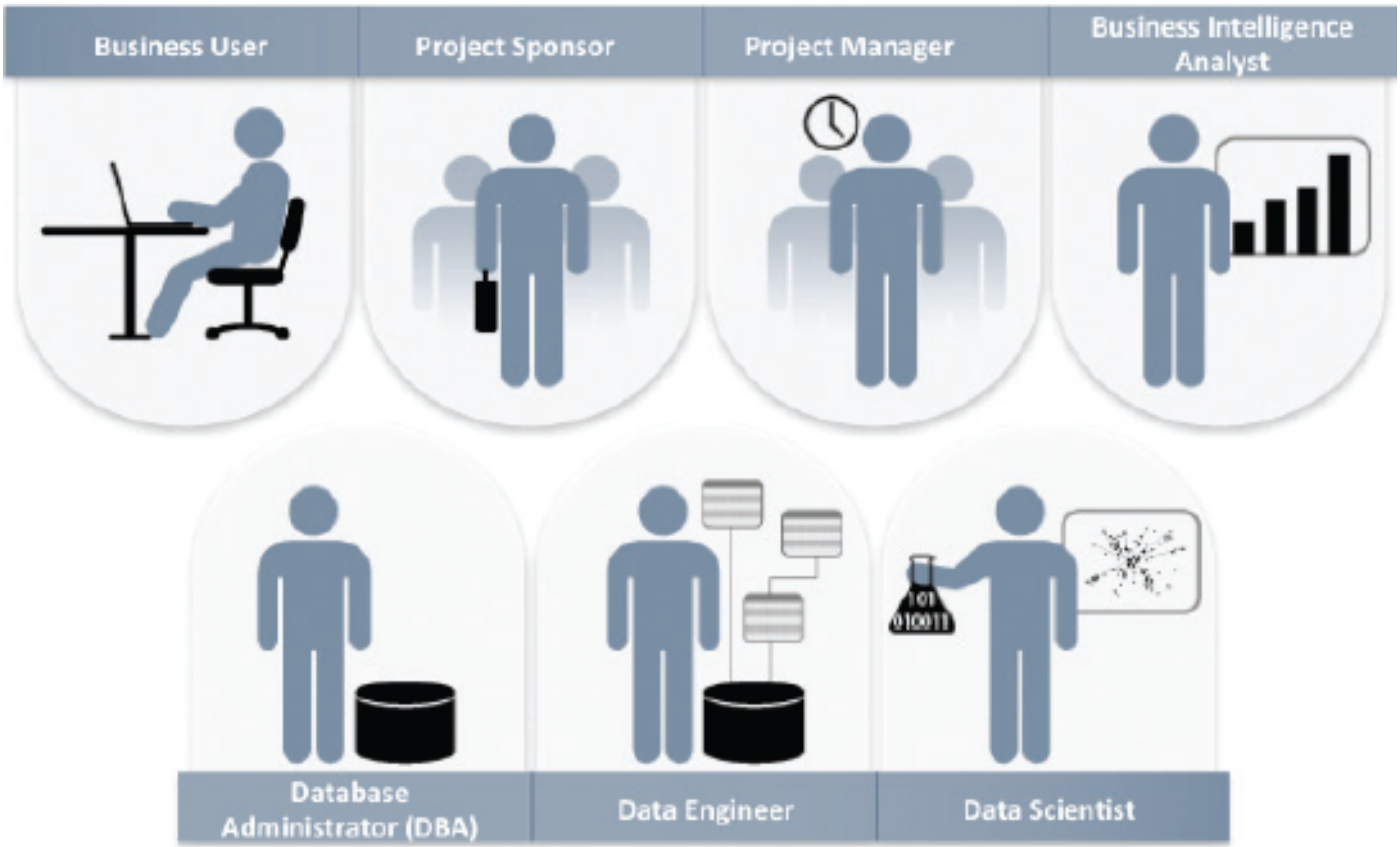
Data Scientist Profile



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

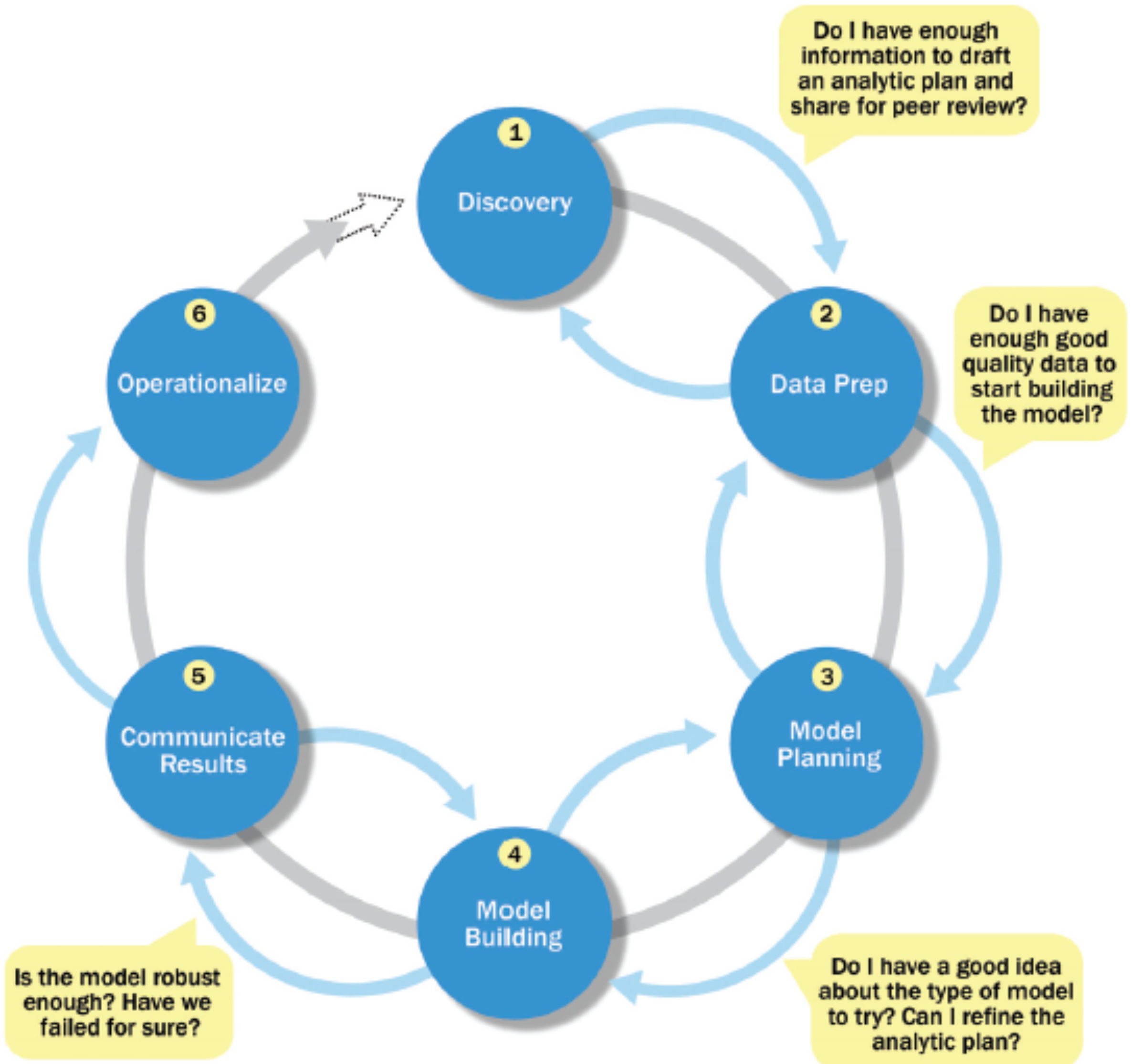
Big Data Analytics Lifecycle

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



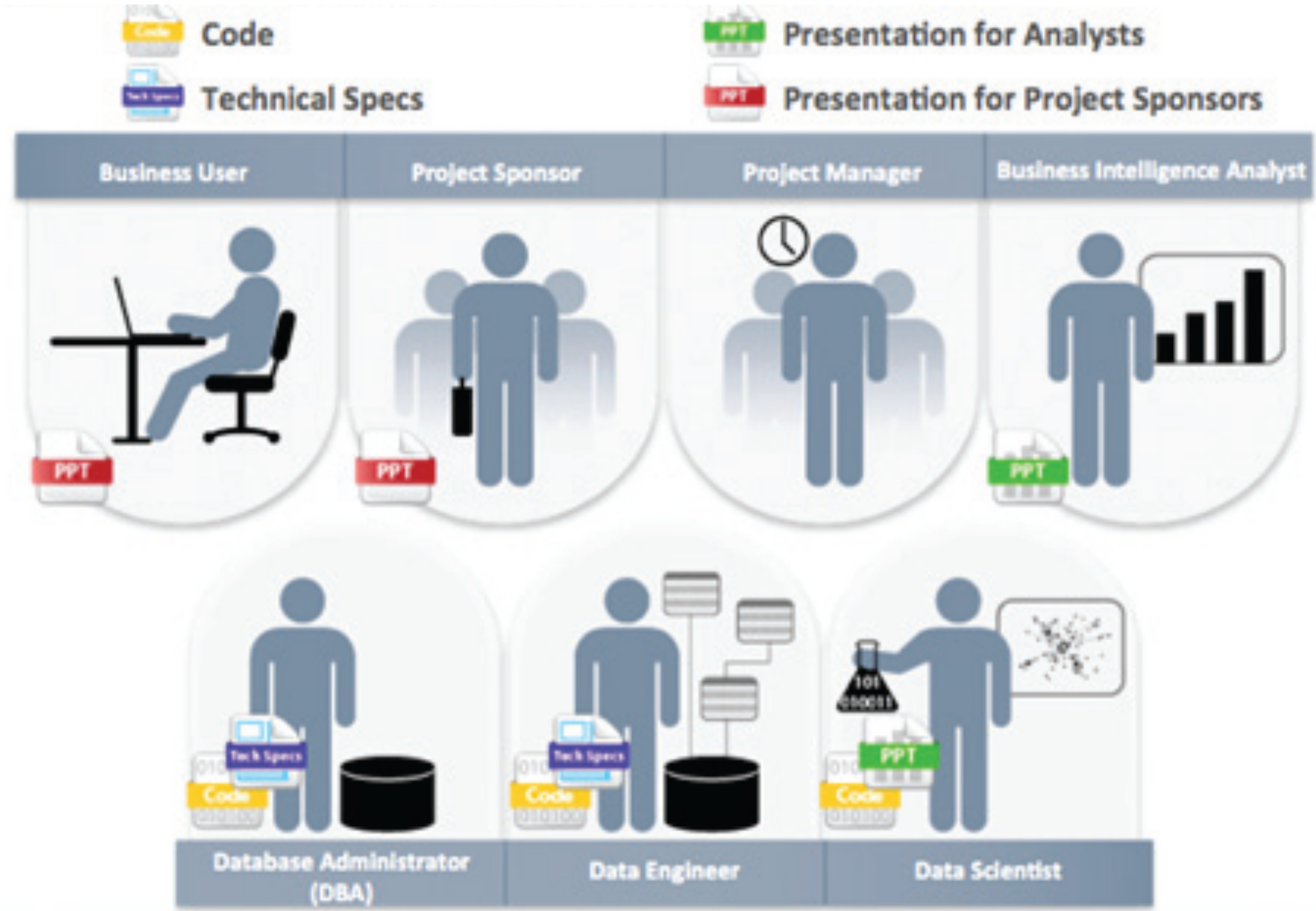
Source: EMC Education Services, Data Science and Big Data Analytics: Discovering, Analyzing, Visualizing and Presenting Data, Wiley, 2015

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)

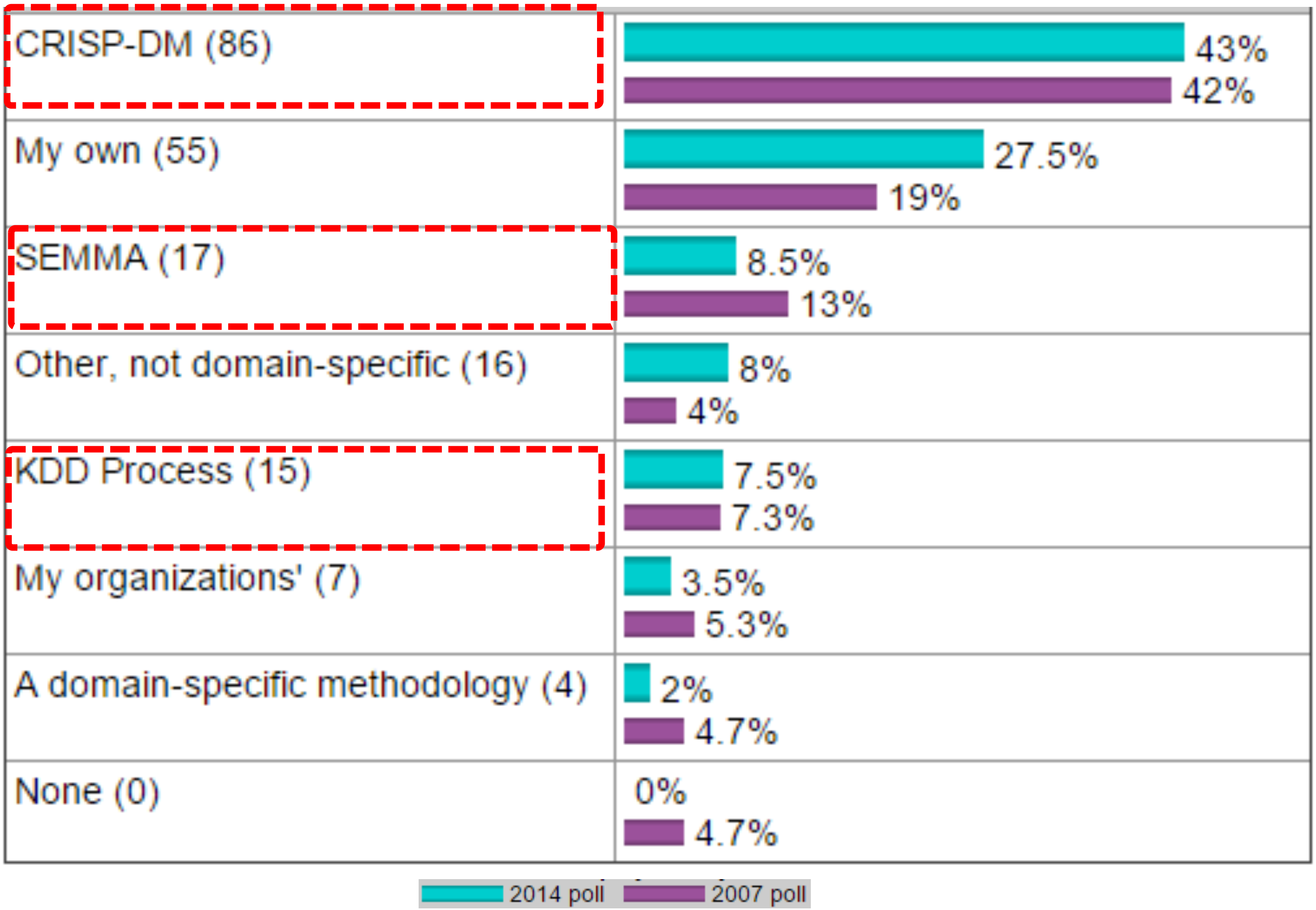
Source: Turban et al. (2011), Decision Support and Business Intelligence Systems

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.

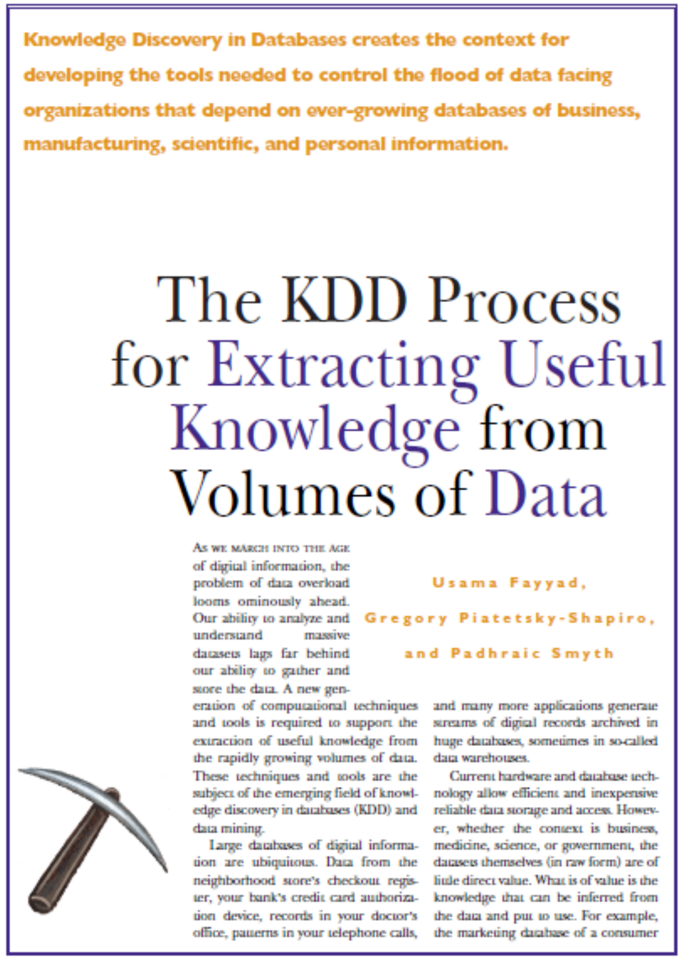
Fayyad, U., Piatetsky-Shapiro, G., & Smyth, P. (1996).

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.

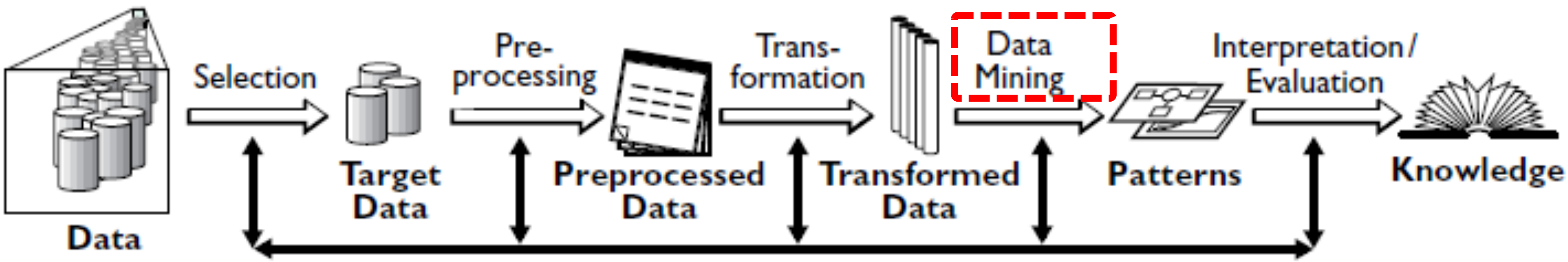
Large databases of digital information are ubiquitous. Data from the neighborhood store's checkout register, your bank's credit card authorization device, records in your doctor's office, patterns in your telephone calls, and many more applications generate streams of digital records archived in huge 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

Data Mining

Knowledge Discovery in Databases (KDD) Process

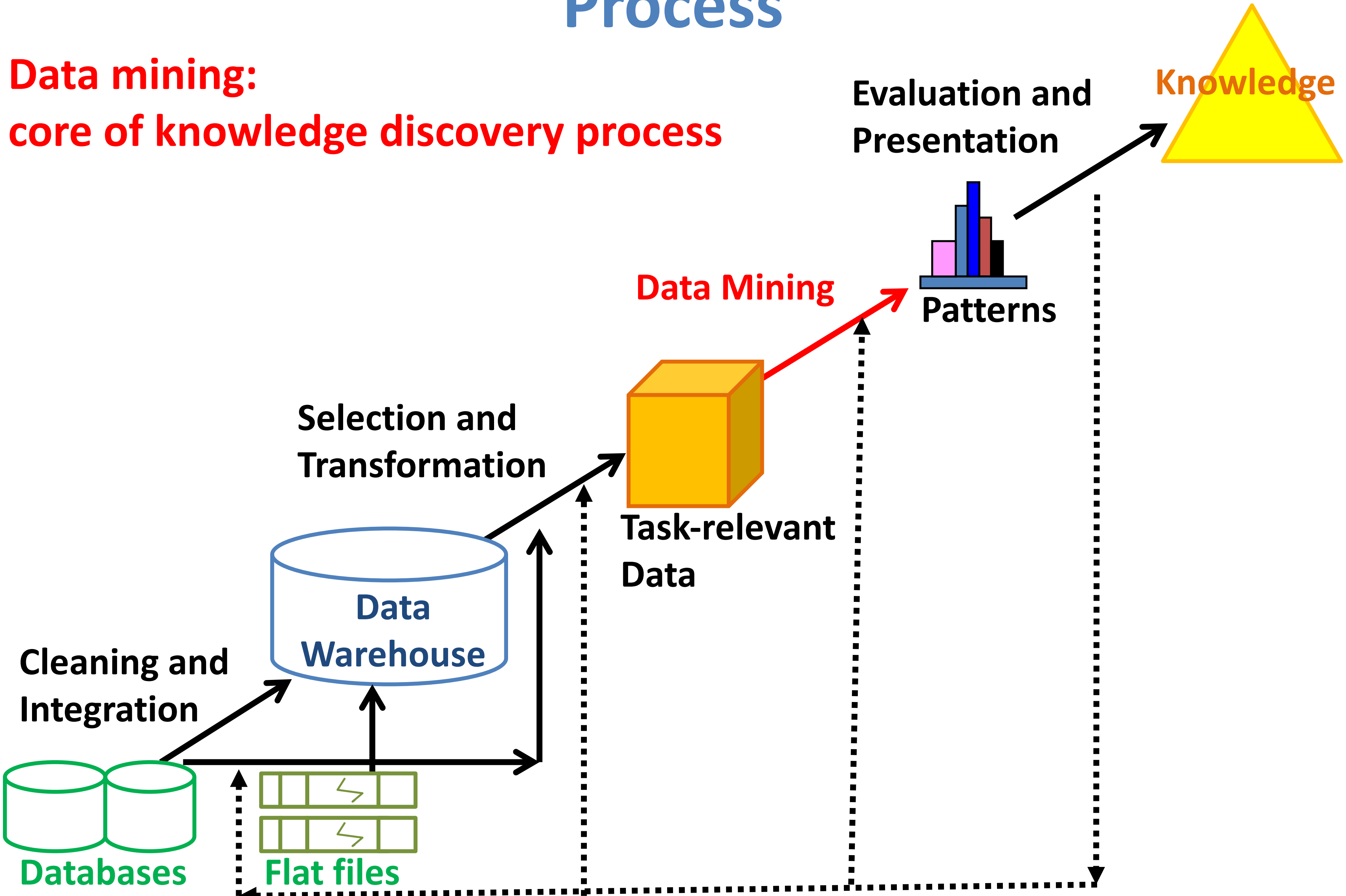
(Fayyad et al., 1996)



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.

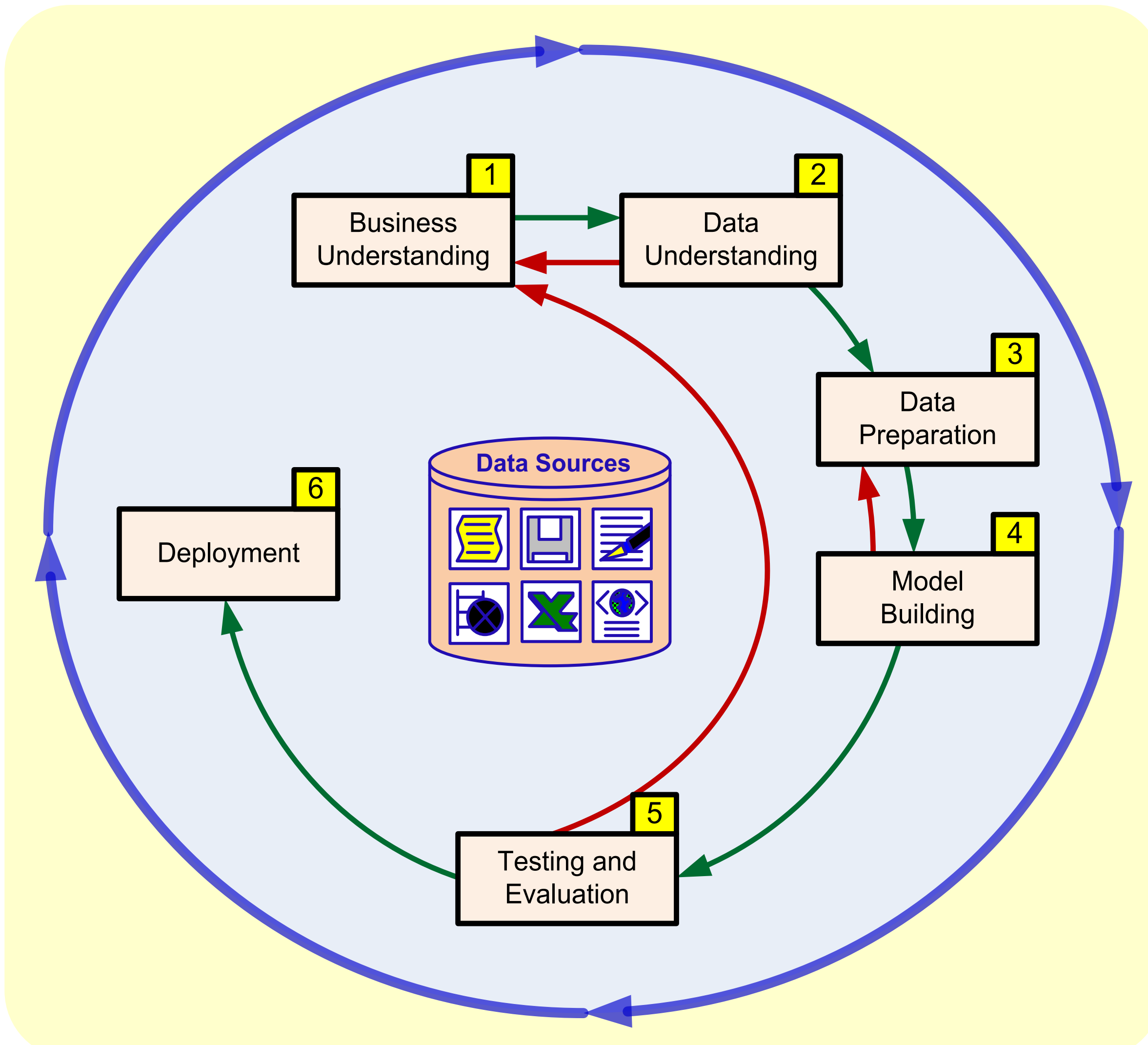
Knowledge Discovery in Databases (KDD) Process

Data mining:
core of knowledge discovery process



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

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

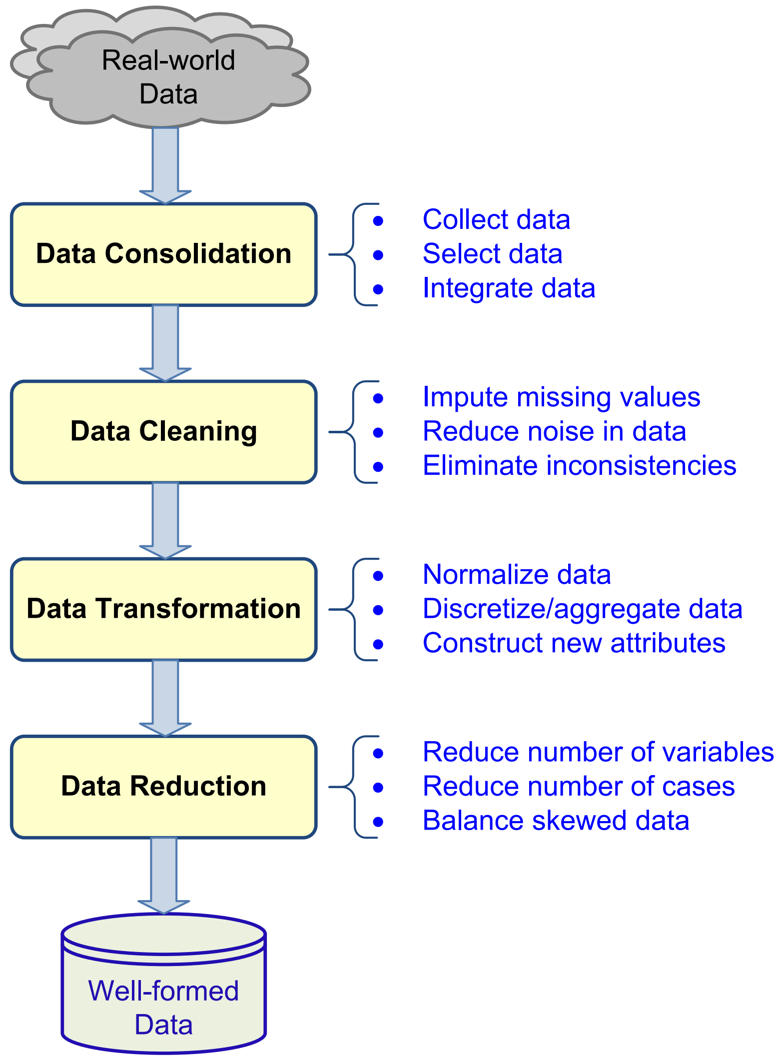
Step 5: Testing and Evaluation

Step 6: Deployment

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

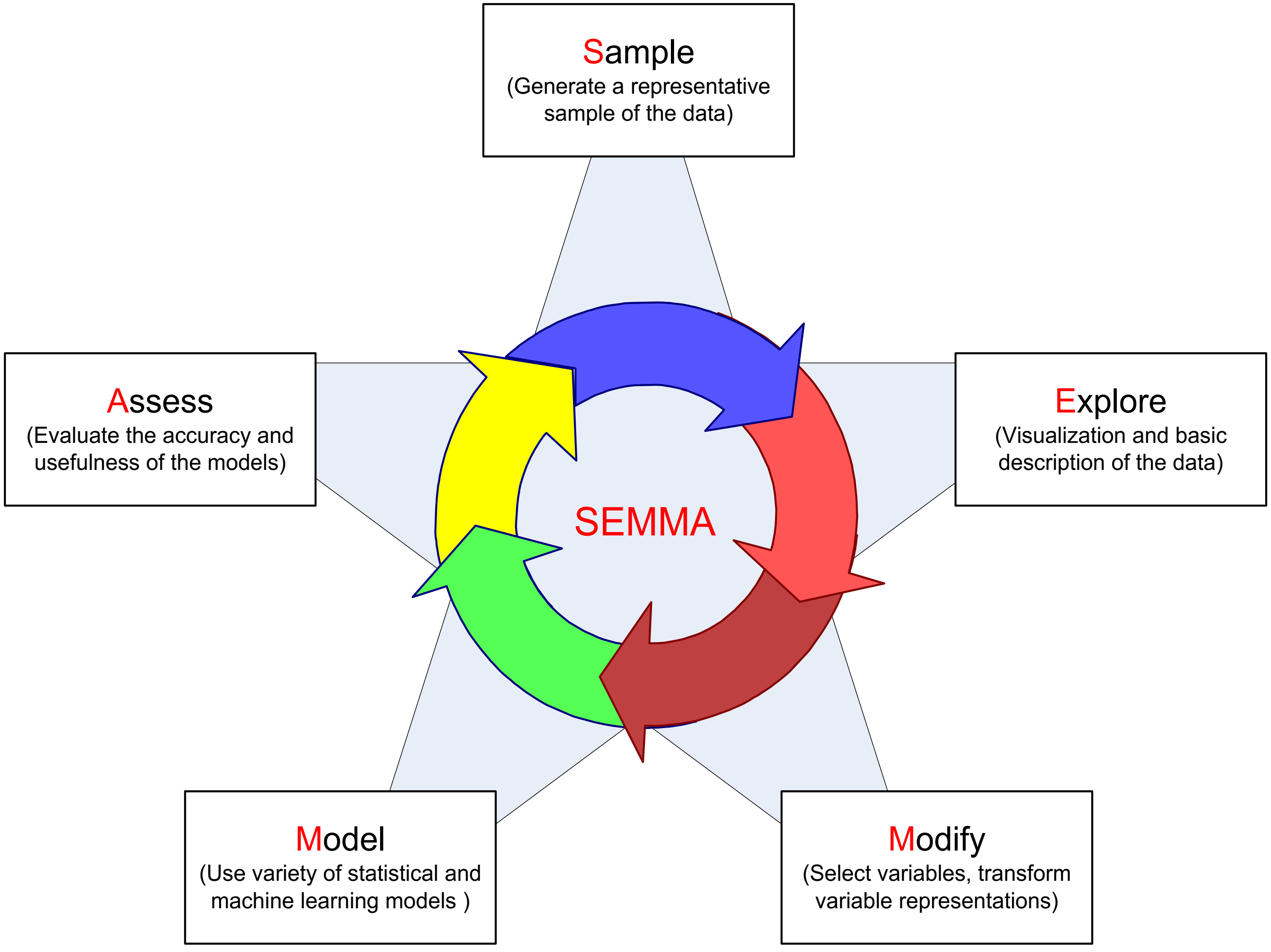
Source: Turban et al. (2011), Decision Support and Business Intelligence Systems

Data Preparation – A Critical DM Task



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

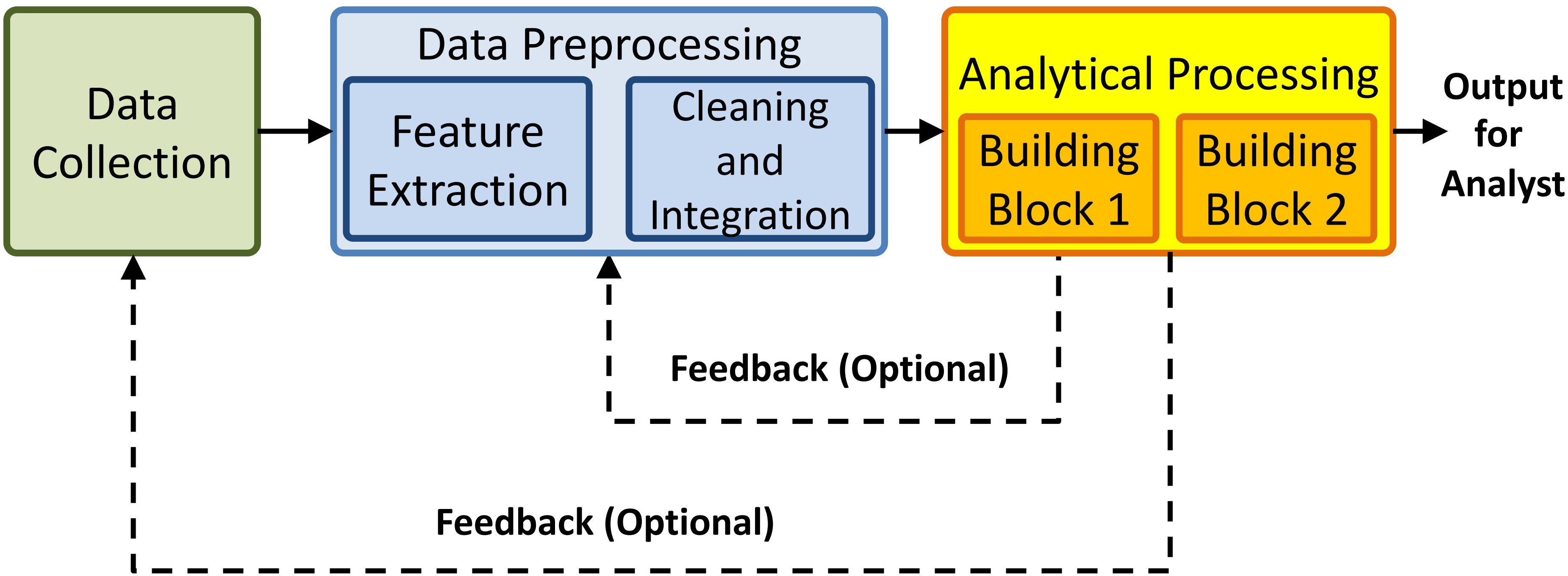
Data Mining Process: SEMMA



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

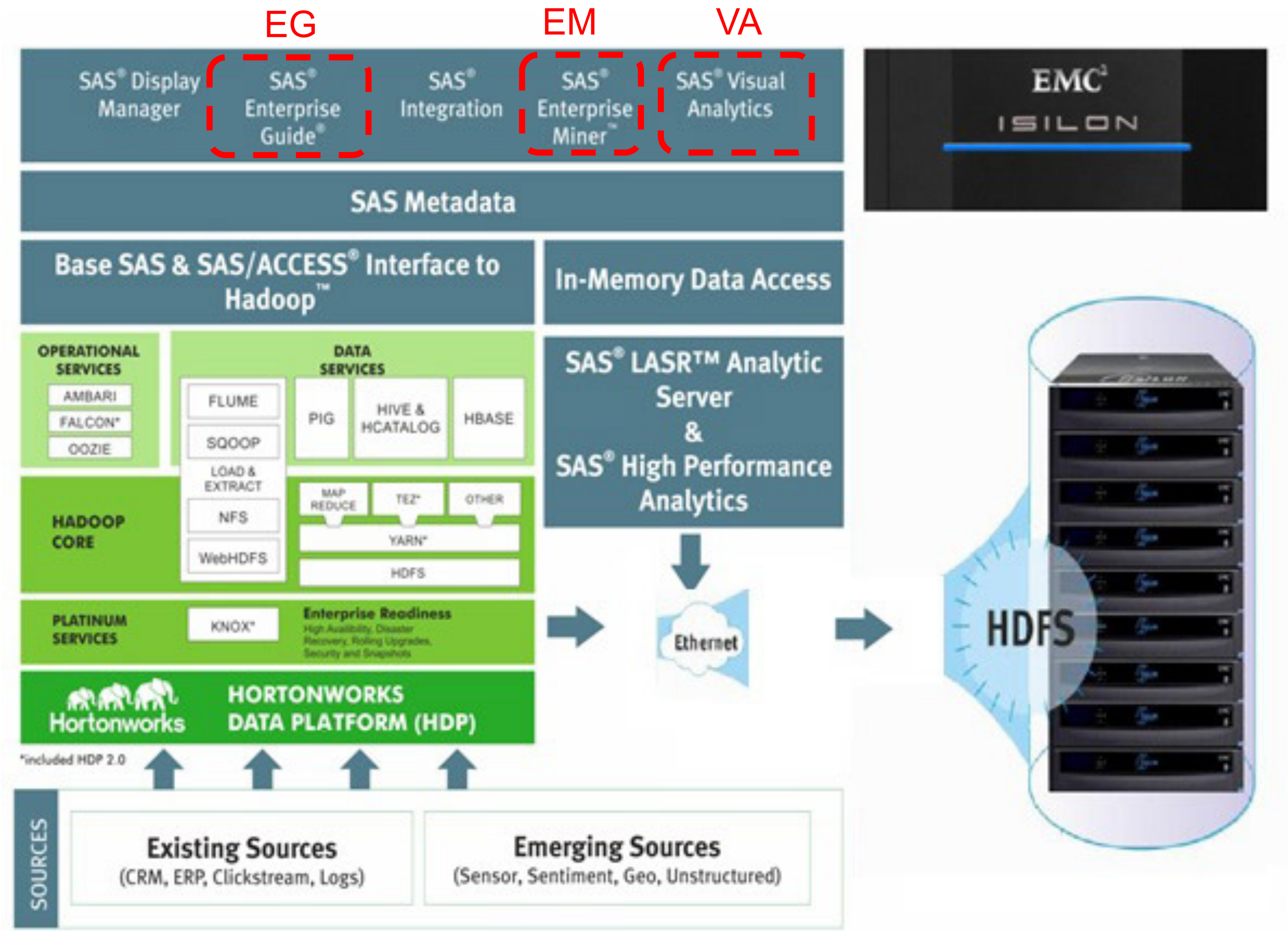
Data Mining Processing Pipeline

(Charu Aggarwal, 2015)

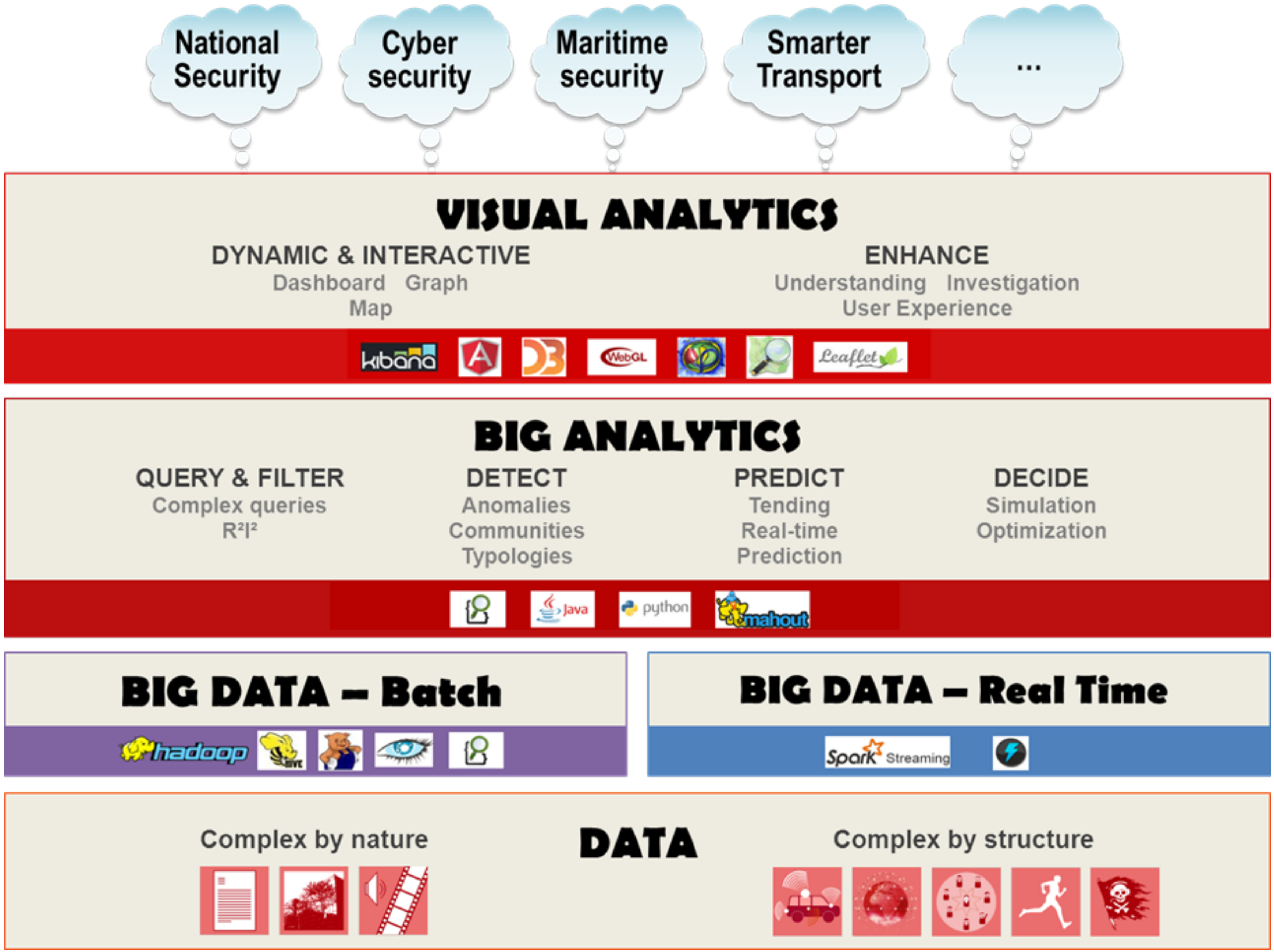


Source: Charu Aggarwal (2015), Data Mining: The Textbook Hardcover, Springer

Big Data Solution



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



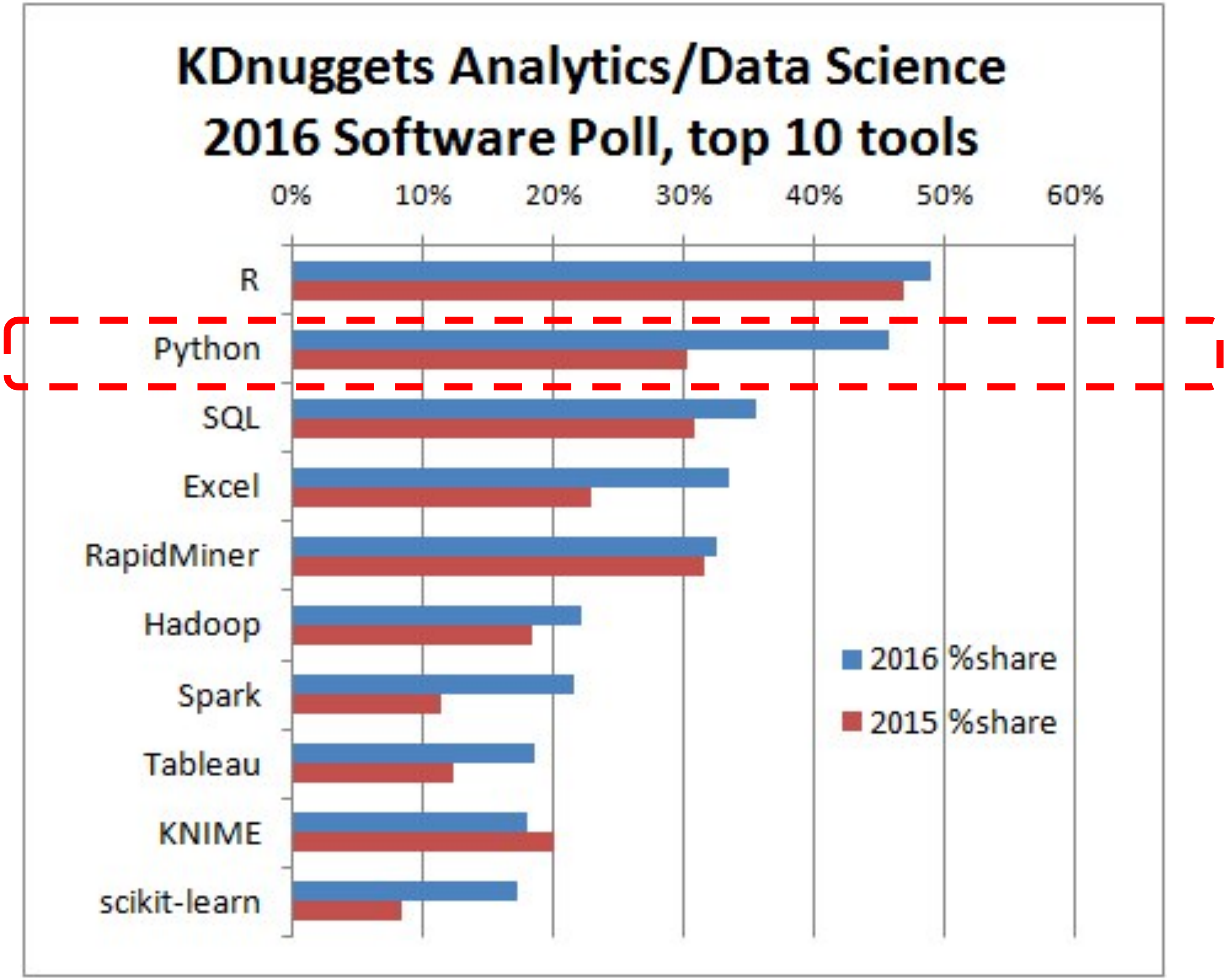
Source: https://www.thalesgroup.com/en/worldwide/big-data/big-data-big-analytics-visual-analytics-what-does-it-all-mean

Python for Big Data Analytics

Language Rank	Types	Spectrum Ranking
1. C		100.0
2. Java		98.1
3. Python		98.0
4. C++		95.9
5. R		87.9
6. C#		86.7
7. PHP		82.8
8. JavaScript		82.2
9. Ruby		74.5
10. Go		71.9

Source: http://spectrum.ieee.org/computing/software/the-2016-top-programming-languages

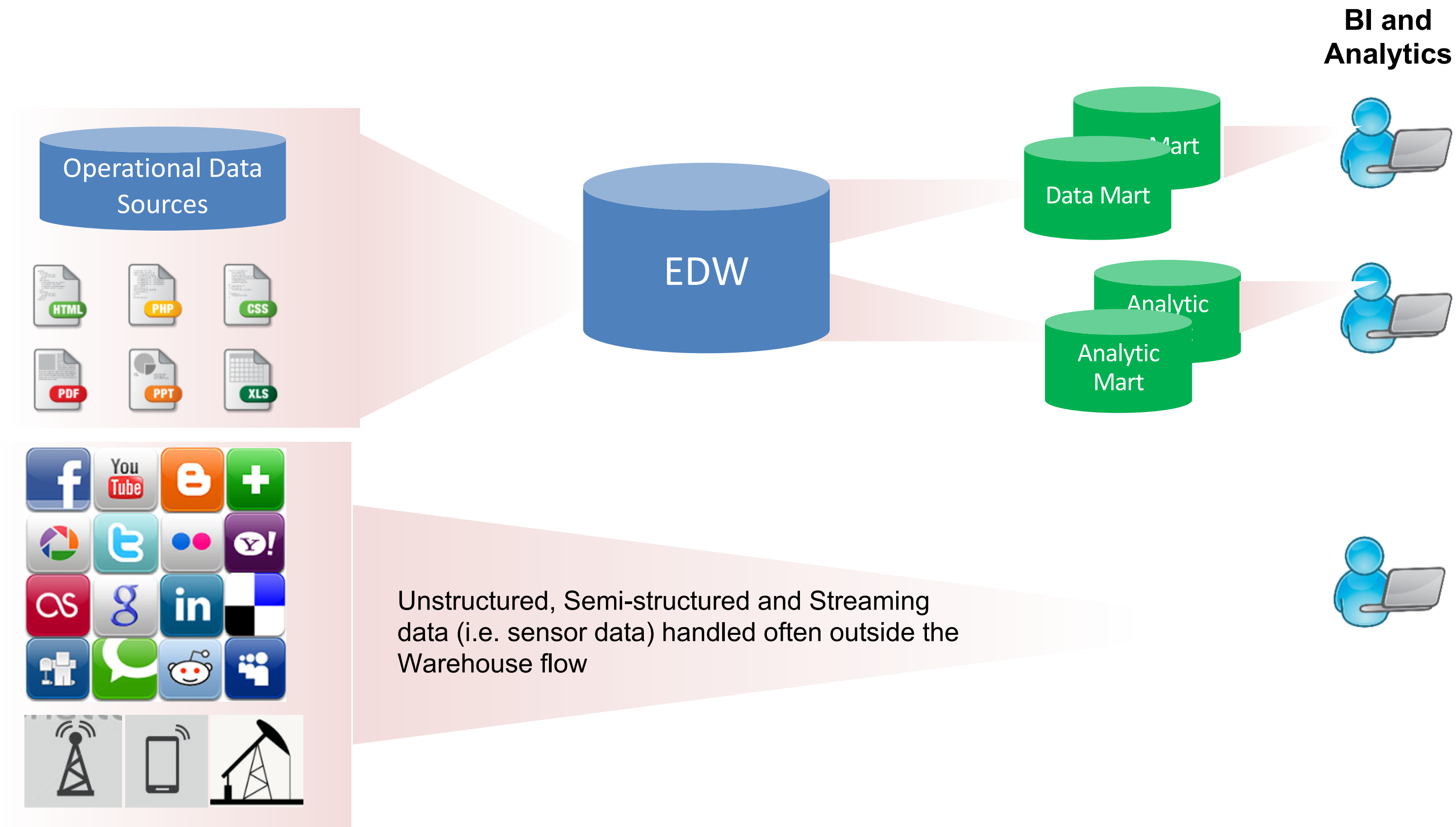
Python: Analytics and Data Science Software



Source: http://www.kdnuggets.com/2016/06/r-python-top-analytics-data-mining-data-science-software.html

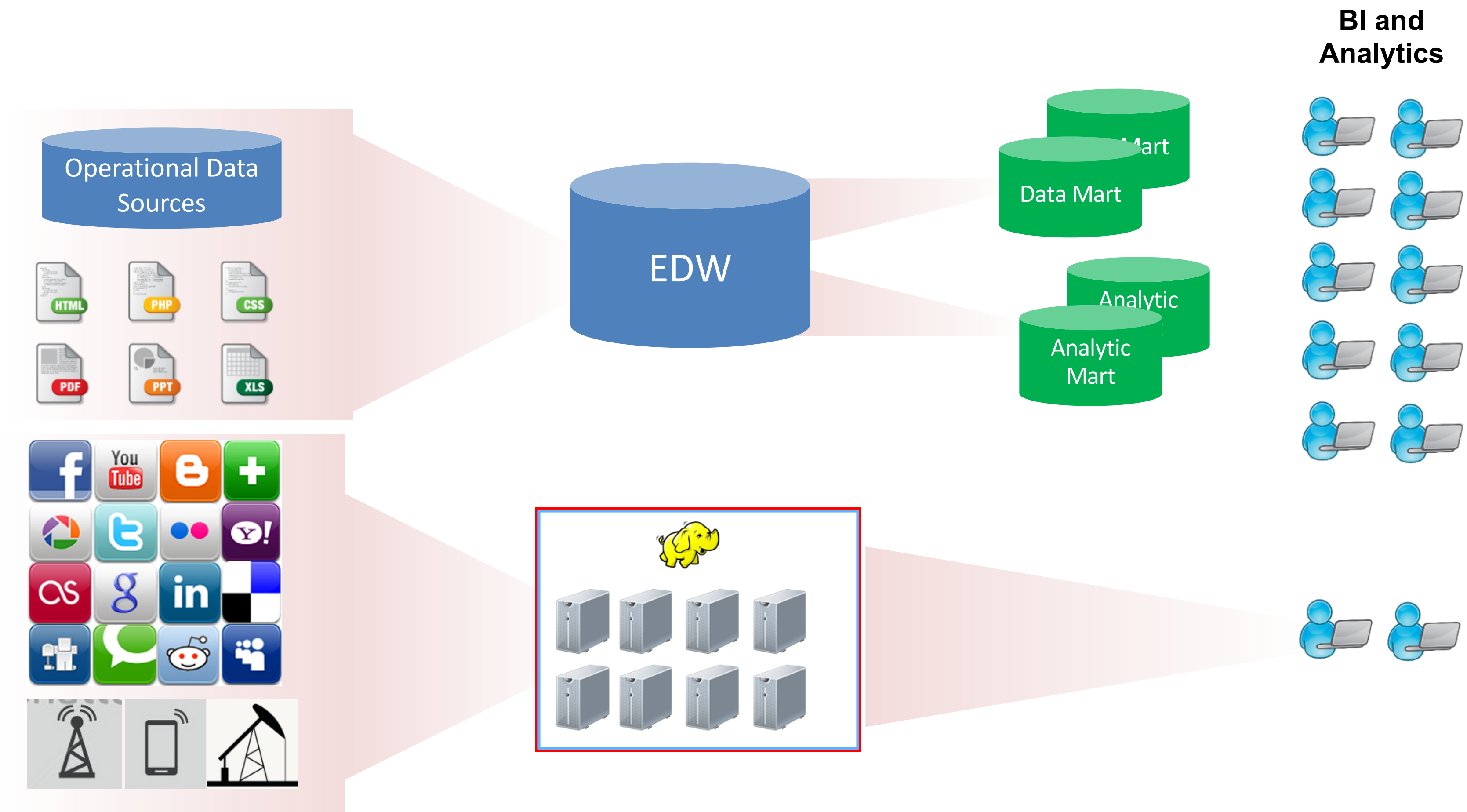
Architectures of Big Data Analytics

Traditional Analytics



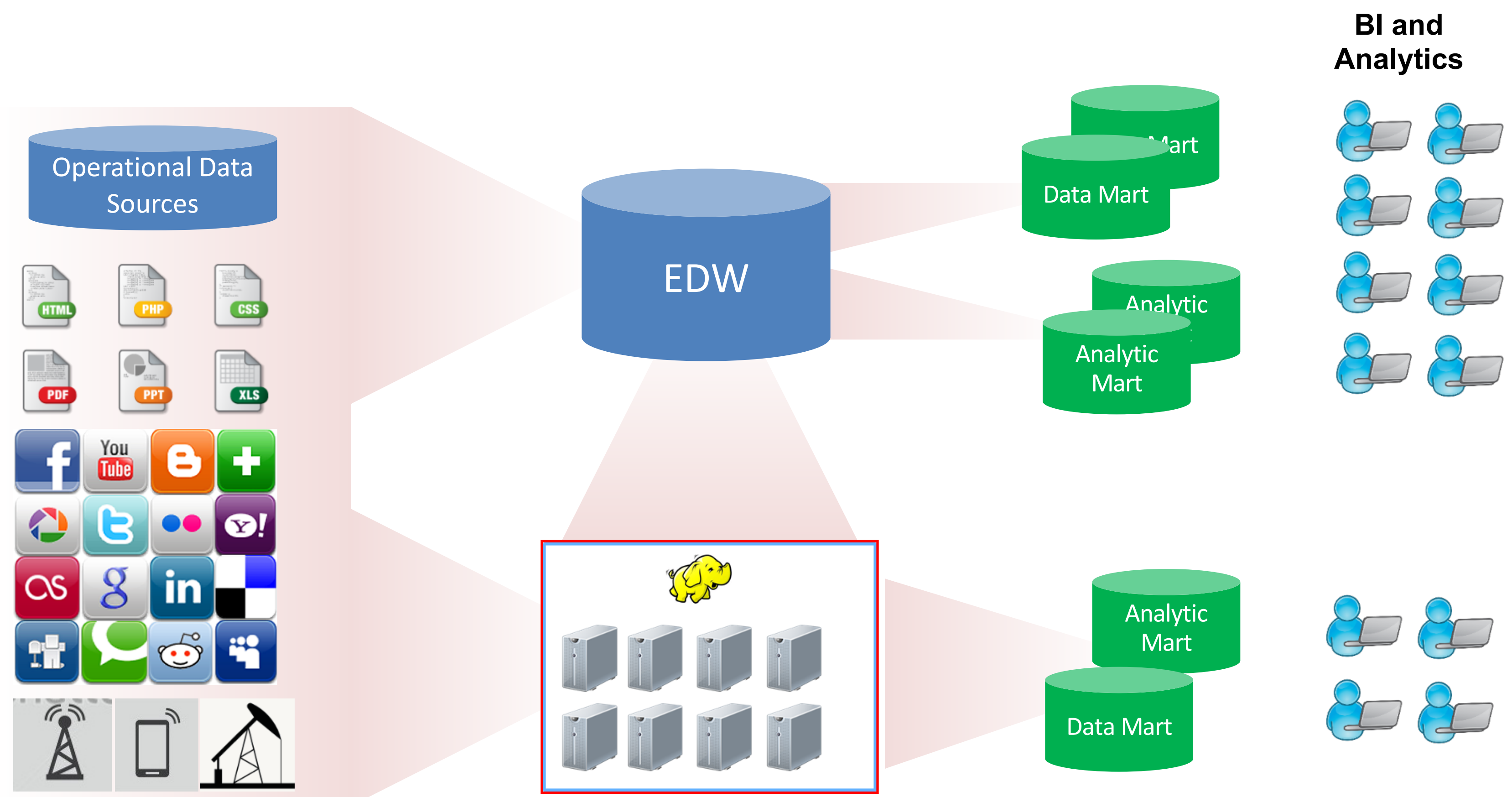
Source: Deepak Ramanathan (2014), SAS Modernization architectures - Big Data Analytics

Hadoop as a "new data" Store



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

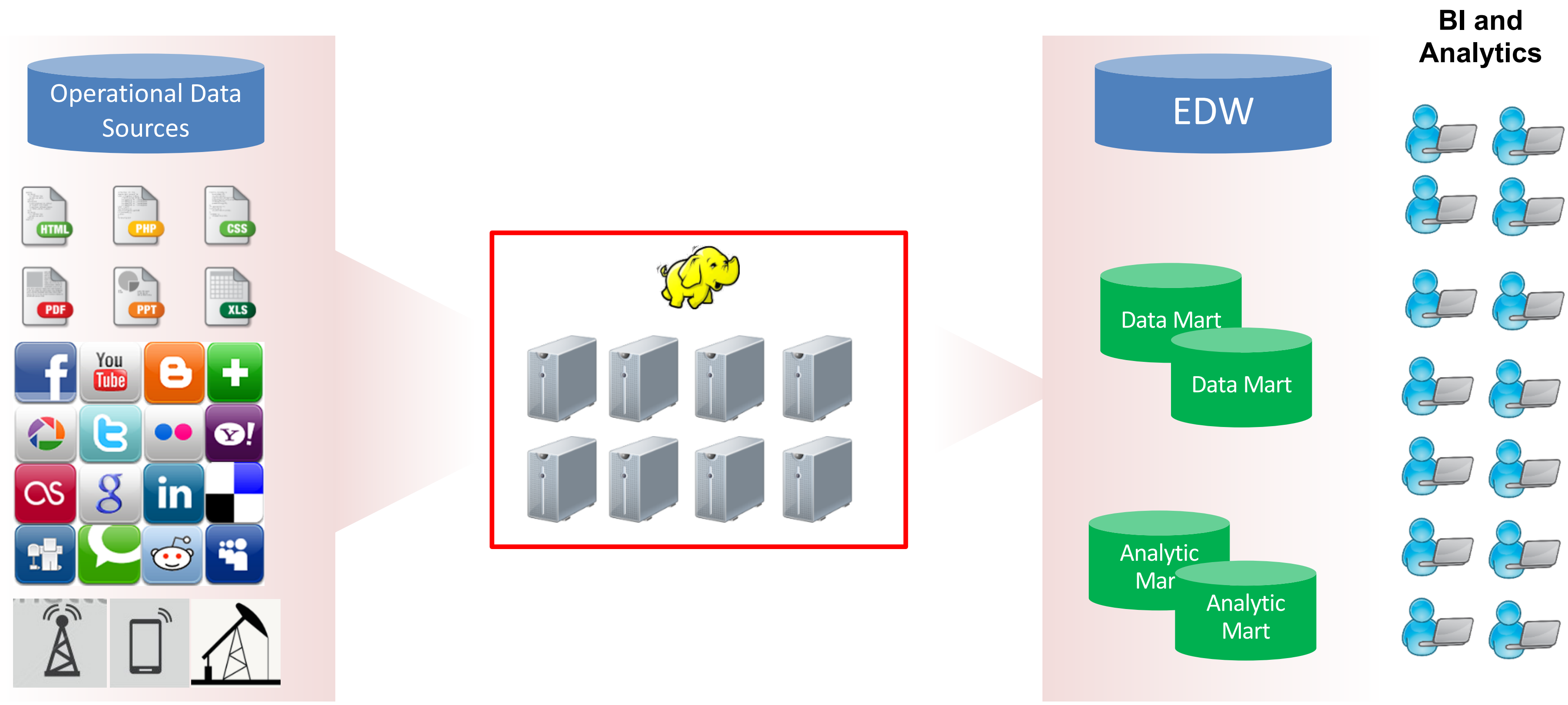
Hadoop as an additional input to the EDW



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

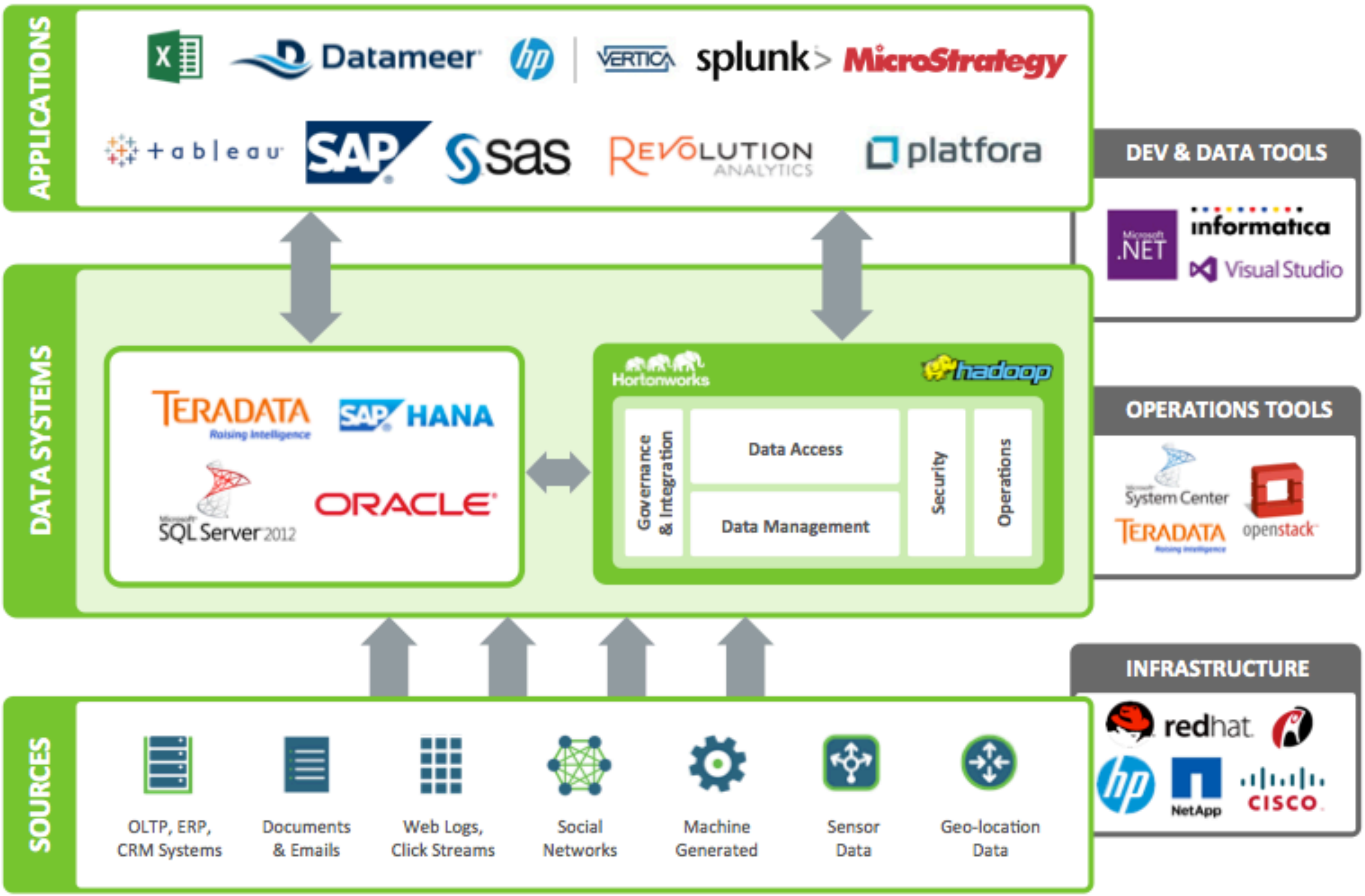
Hadoop Data Platform As a "staging Layer" as part of a "data Lake"

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



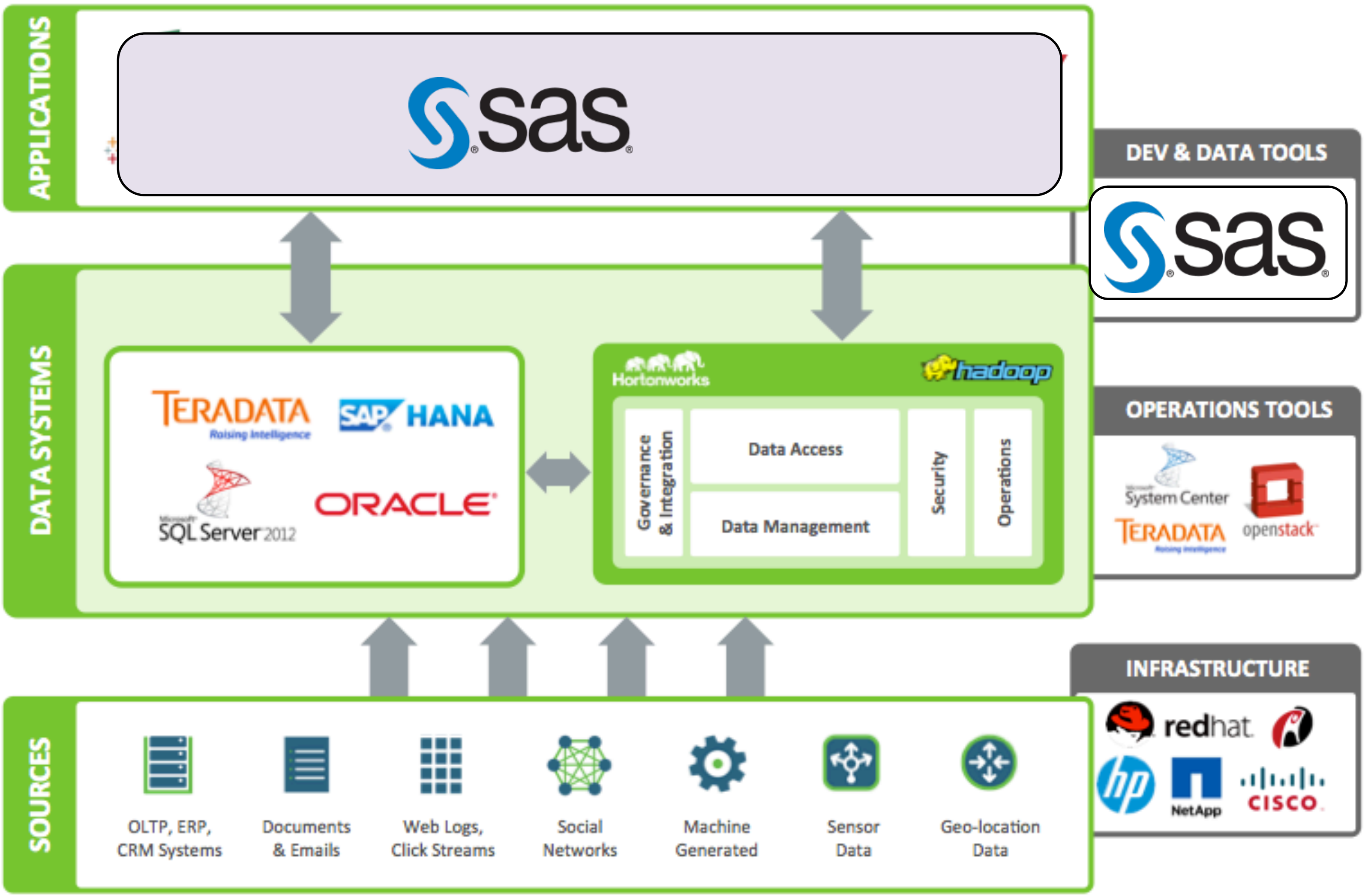
Source: Deepak Ramanathan (2014), SAS Modernization architectures - Big Data Analytics

SAS Big data Strategy – SAS areas



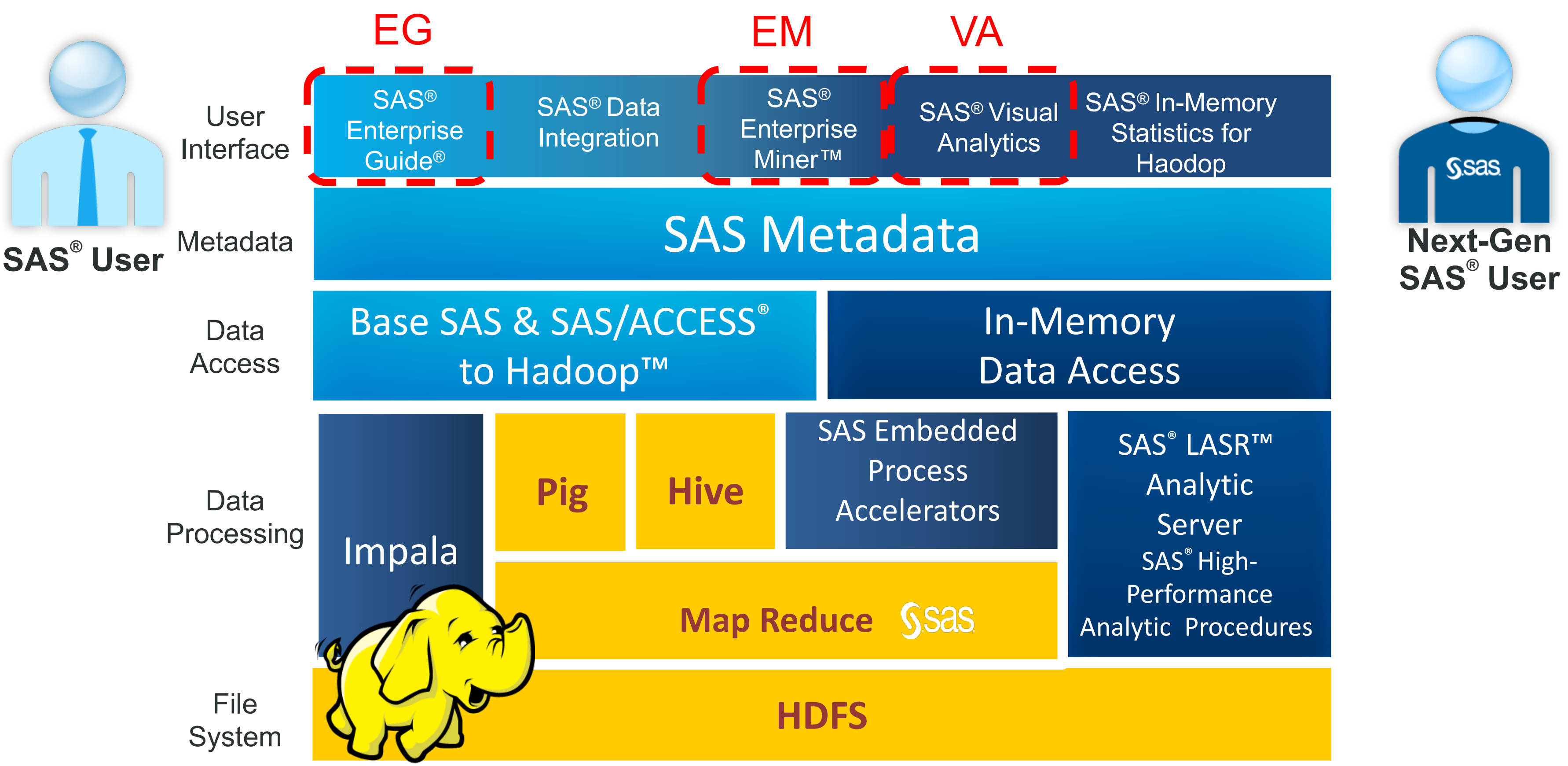
Source: Deepak Ramanathan (2014), SAS Modernization architectures - Big Data Analytics

SAS Big data Strategy – SAS areas



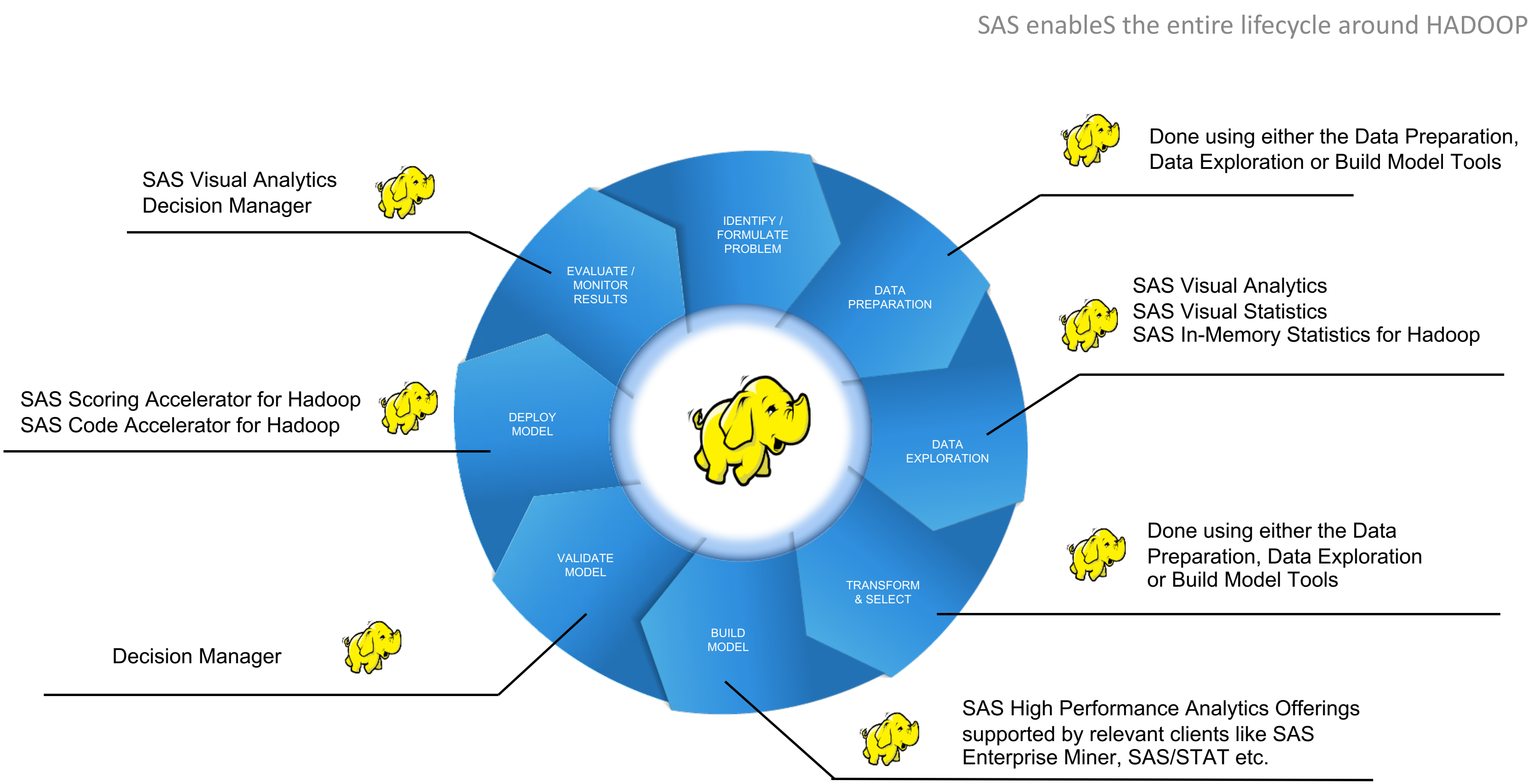
Source: Deepak Ramanathan (2014), SAS Modernization architectures - Big Data Analytics

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

SAS® VISUAL ANALYTICS

A Single solution for Data Discovery, Visualization, analytics and reporting

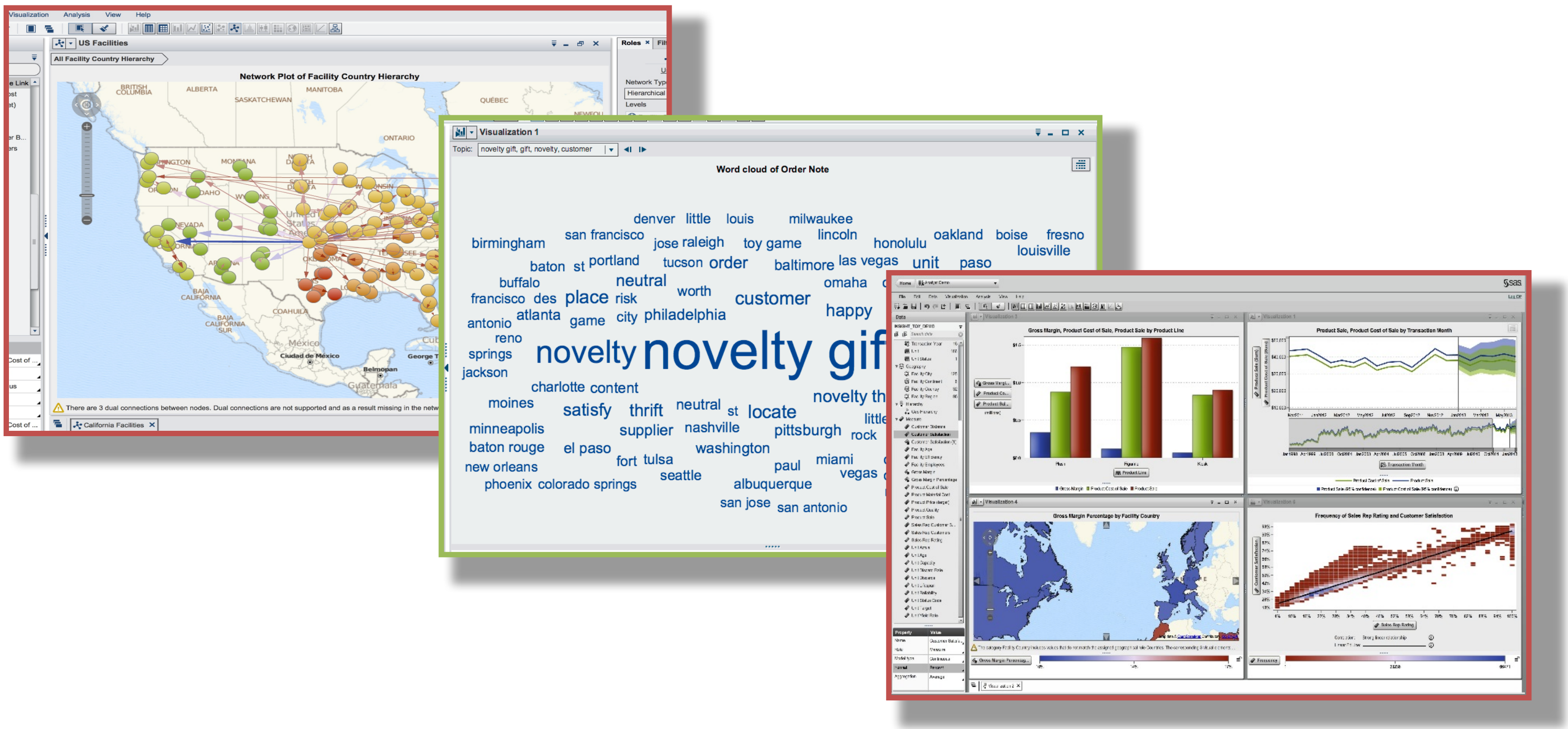
Source: Deepak Ramanathan (2014), SAS Modernization architectures - Big Data Analytics

SAS® VISUAL ANALYTICS

Example: text analysis gives you insight to customer experience and opinion

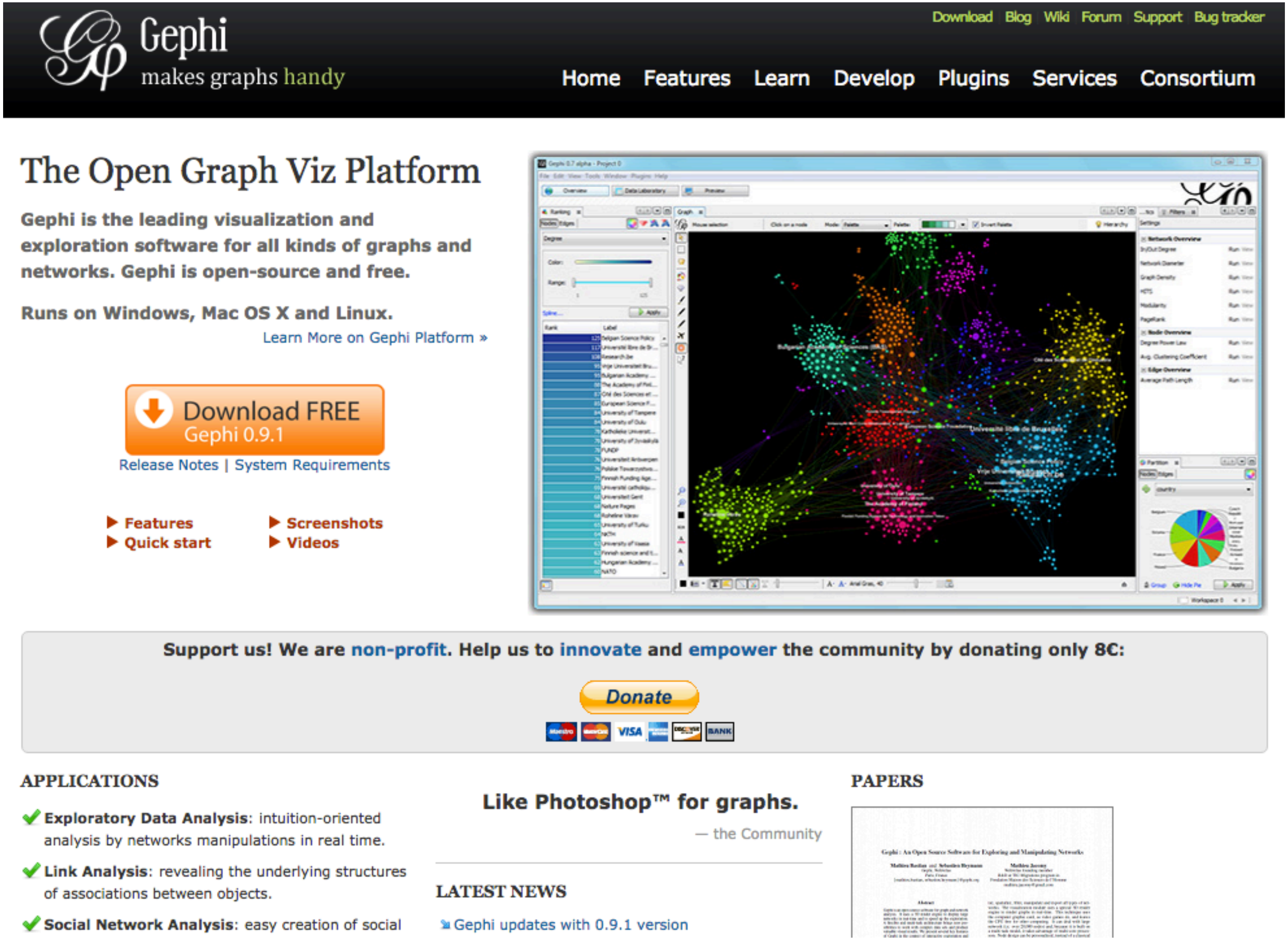
Analytics applied to text provides real MEANING

Visualization



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

Gephi



https://gephi.org/

igraph

http://igraph.org/redirect.html

sigma.js

σsigmajs

GET STARTED FEATURES USE CASES TUTORIAL REFERENCES

sigmajs

TUTORIAL v1.1.0 DOWNLOAD

Sigma is a JavaScript library **dedicated to graph drawing**. It makes easy to publish networks on Web pages, and allows developers to integrate network exploration in rich Web applications.

FORK ON GITHUB FOLLOW ON TWITTER DOWNLOAD V1.1.0 SCIENCESPO - MÉDIALAB

http://sigmajs.org/

D3.js

Overview Examples Documentation Source

Fork me on GitHub

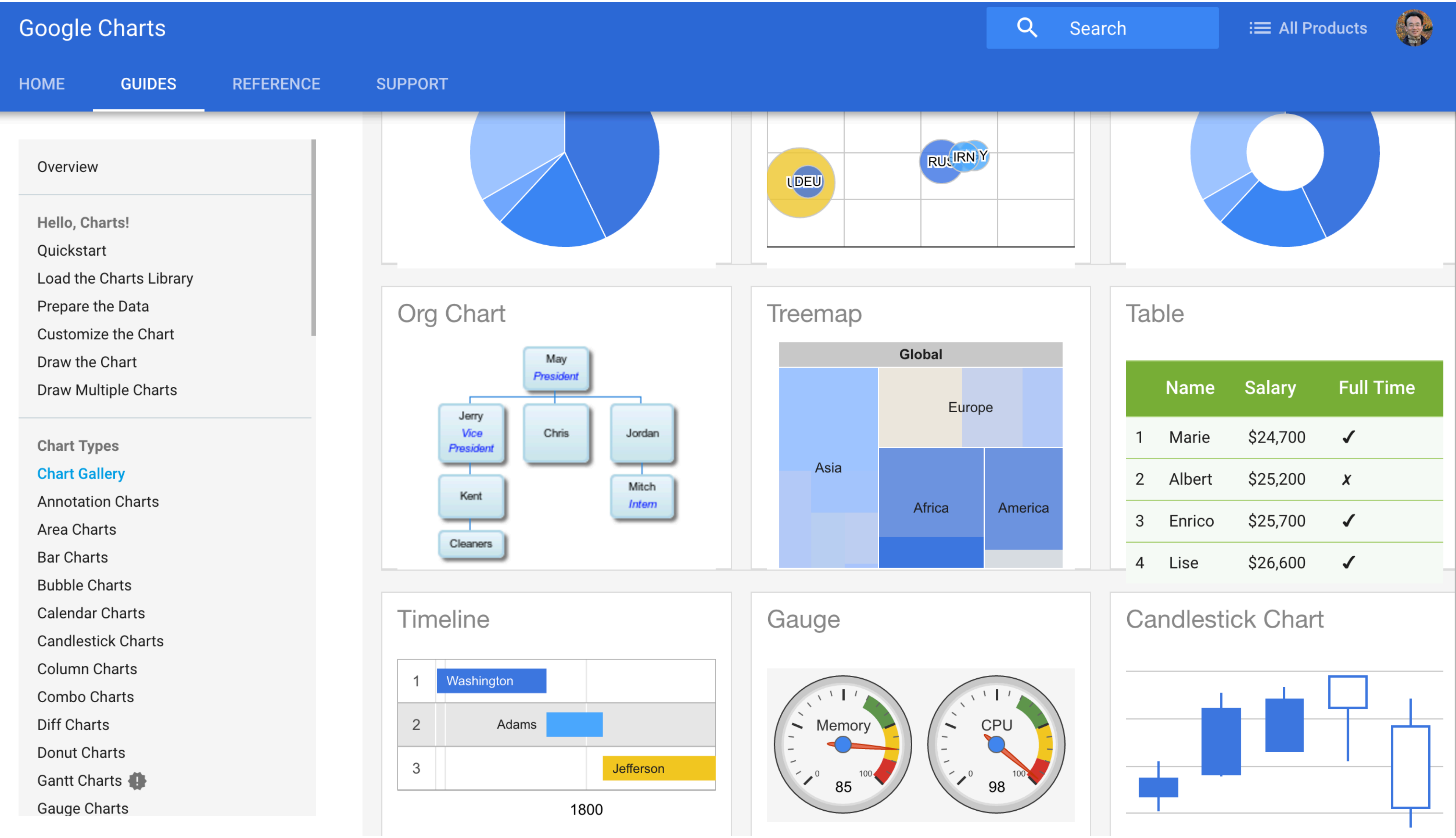
Data-Driven Documents

D3.js is a JavaScript library for manipulating documents based on data. **D3** helps you bring data to life using HTML, SVG, and CSS. D3's emphasis on web standards gives you the full capabilities of

See more examples.

https://d3js.org/

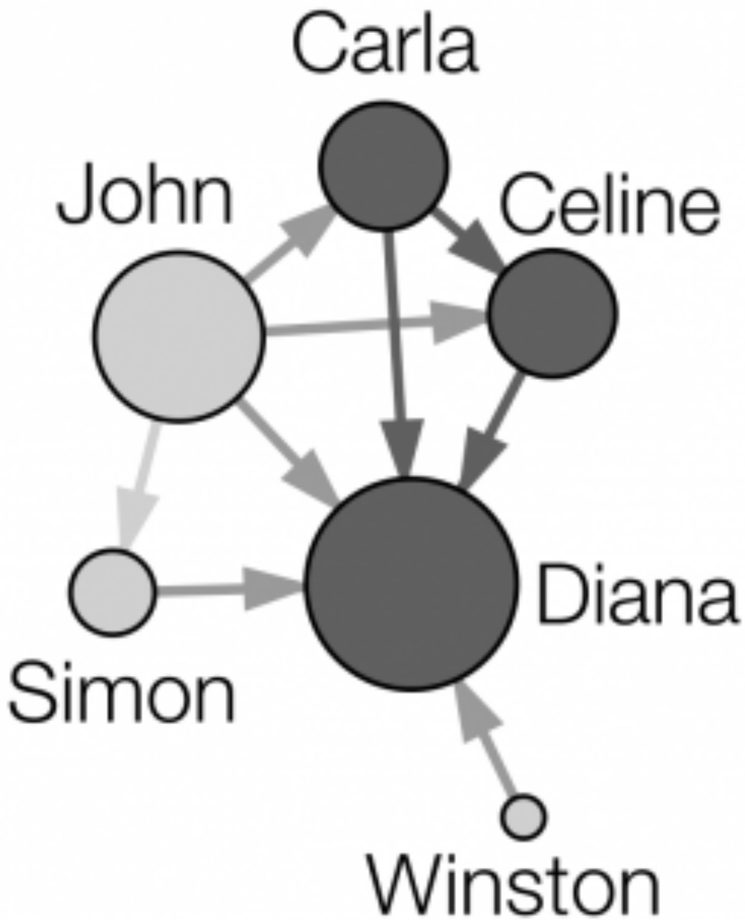
Google Charts



https://developers.google.com/chart/interactive/docs/gallery

Gephi:
Social Network Analysis and Visualization

Network Analysis and Visualization with Gephi



Nodes

```
Id,Label,Attribute
1,John,1
2,Carla,2
3,Simon,1
4,Celine,2
5,Winston,1
6,Diana,2
```

Edges

```
Source,Target
1,2
1,3
1,4
1,6
2,4
2,6
3,6
4,6
5,6
```

Source: http://www.martingrandjean.ch/gephi-introduction/

Nodes and Edges
CSV Text Data for Gephi

Nodes1.csv

```
Id,Label,Attribute
1,John,1
2,Carla,2
3,Simon,1
4,Celine,2
5,Winston,1
6,Diana,2
```

Edges1.csv

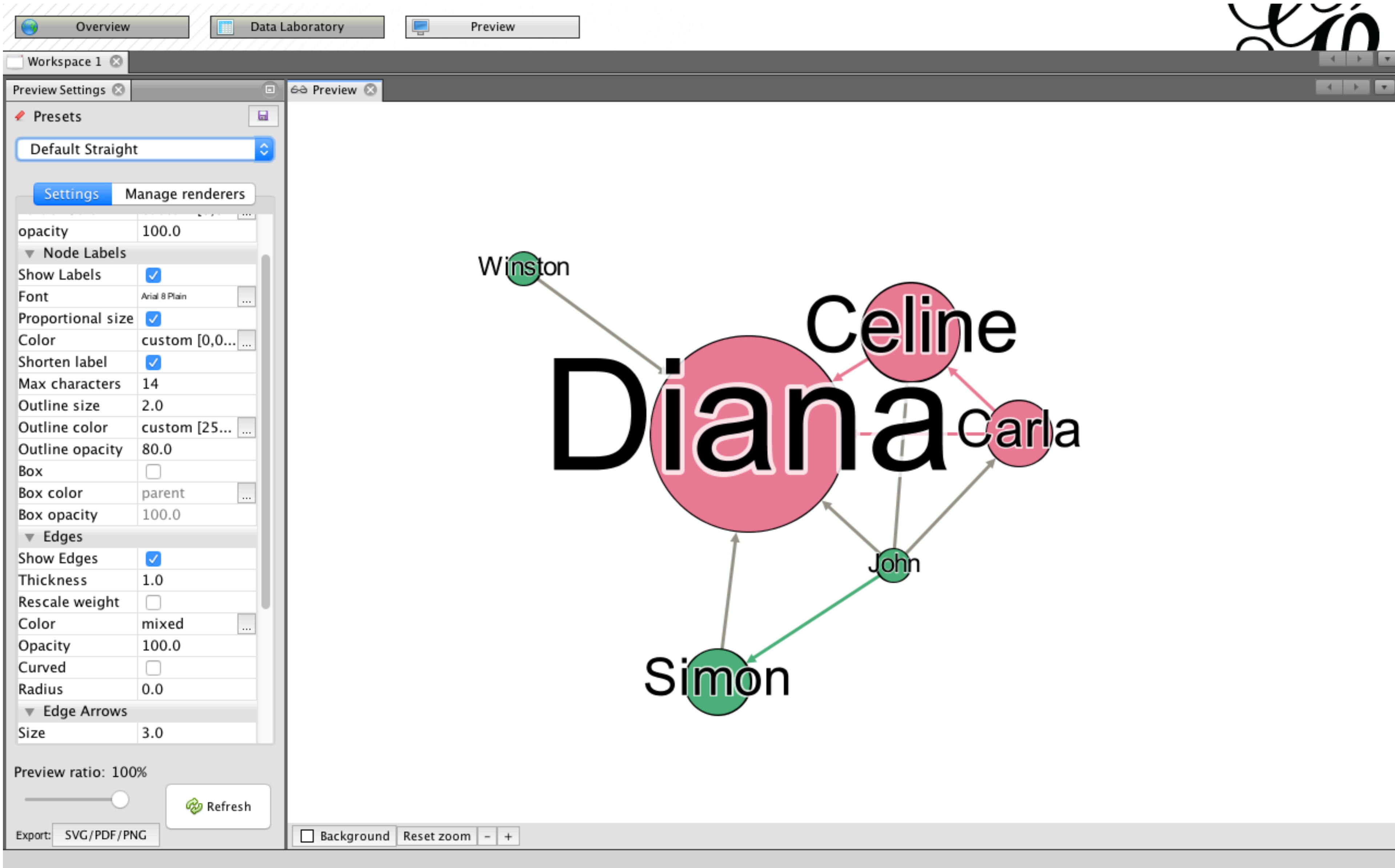
```
Source,Target
1,2
1,3
1,4
1,6
2,4
2,6
3,6
4,6
5,6
```

Nodes1.csv ×


```
Id,Label,Attribute
1,John,1
2,Carla,2
3,Simon,1
4,Celine,2
5,Winston,1
6,Diana,2
```


Source: http://www.martingrandjean.ch/gephi-introduction/

Network Visualization with Gephi



References

- EMC Education Services (2015), Data Science and Big Data Analytics: Discovering, Analyzing, Visualizing and Presenting Data, Wiley
- SAS Modernization architectures - Big Data Analytics, http://www.slideshare.net/deepakramanathan/sas-modernization-architectures-big-data-analytics