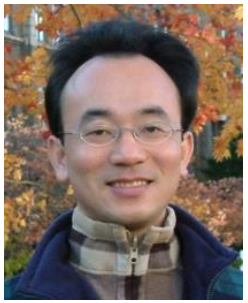
Tamkang University
淡江大學

Tamkang University

李御璽 教授
銘傳大學資訊工程學系

Big Data Analytics on Social Media
(社群媒體大數據分析)

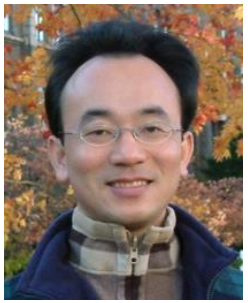
Time: 2015/12/25 (14:00-15:30)
Place: S402, Ming Chuan University



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/

2015-12-25



戴敏育 博士
(Min-Yuh Day, Ph.D.)

淡江大學資管系專任助理教授

中央研究院資訊科學研究所訪問學人

國立台灣大學資訊管理博士

Publications Co-Chairs, IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM 2013-)

Program Co-Chair, IEEE International Workshop on Empirical Methods for Recognizing Inference in TExt (IEEE EM-RITE 2012-)

Workshop Chair, The IEEE International Conference on Information Reuse and Integration (IEEE IRI)

Outline

- Big Data Analytics on Social Media
- Analyzing the Social Web:
  Social Network Analysis
- NTCIR 12 QALab-2 Task

Social Media

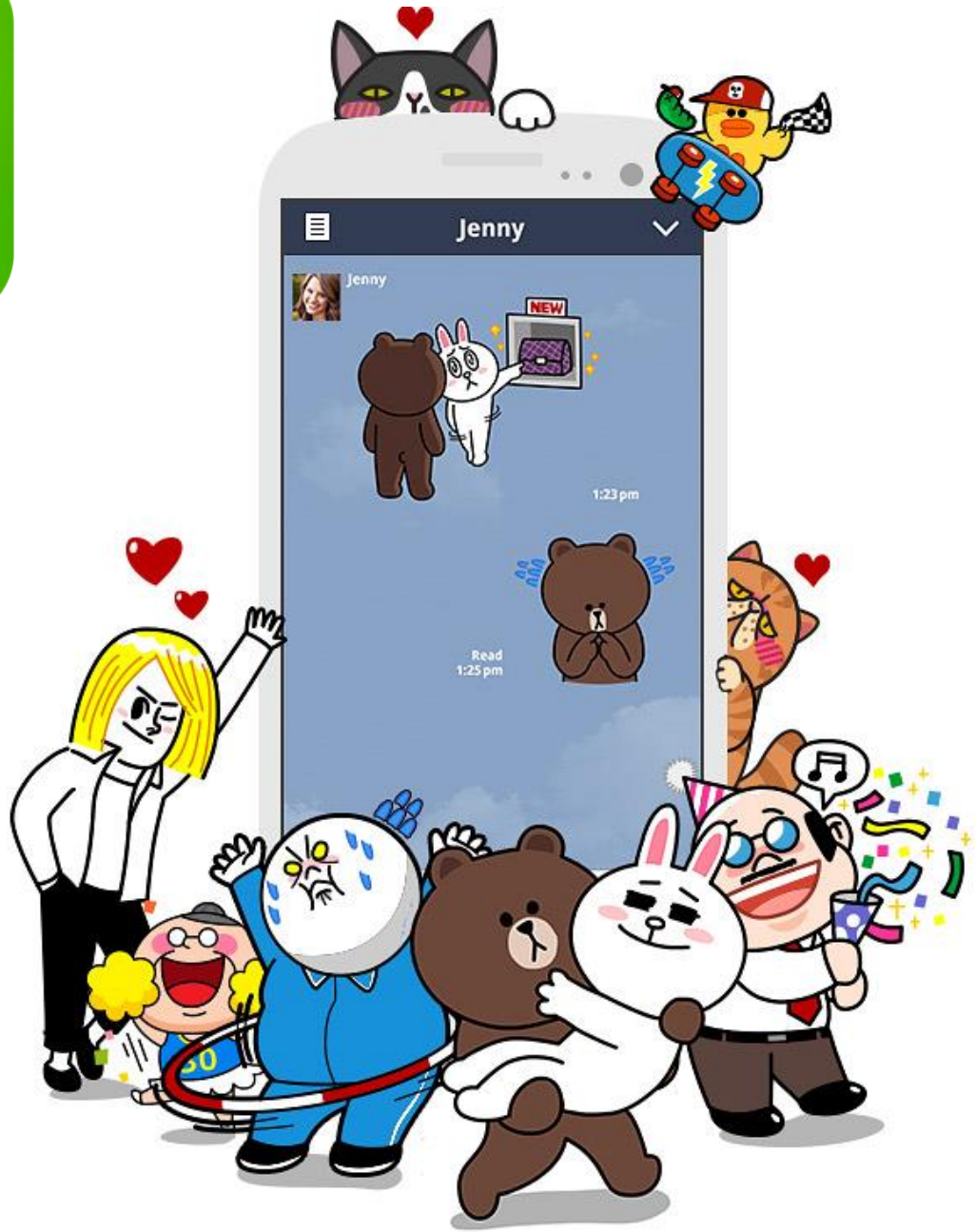
Source: http://www.dreamstime.com/royalty-free-stock-images-christmas-tree-social-media-icons-image21457239

Social Media

Source: http://hungrywolfmarketing.com/2013/09/09/what-are-your-social-marketing-goals/

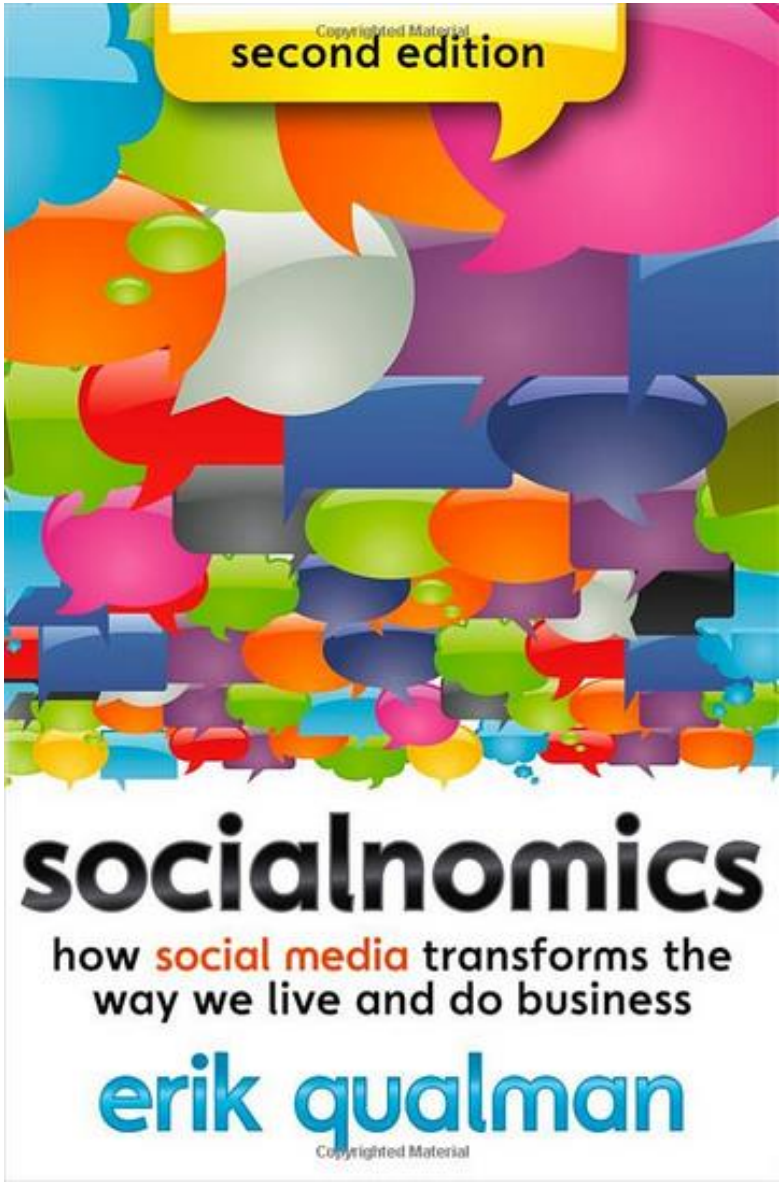
Social Media

Source: http://blog.contentfrog.com/wp-content/uploads/2012/09/New-Social-Media-Icons.jpg



Source: http://line.me/en/

Socialnomics



Source: http://www.amazon.com/Socialnomics-Social-Media-Transforms-Business/dp/1118232658

Emotions

Love

Anger

Joy

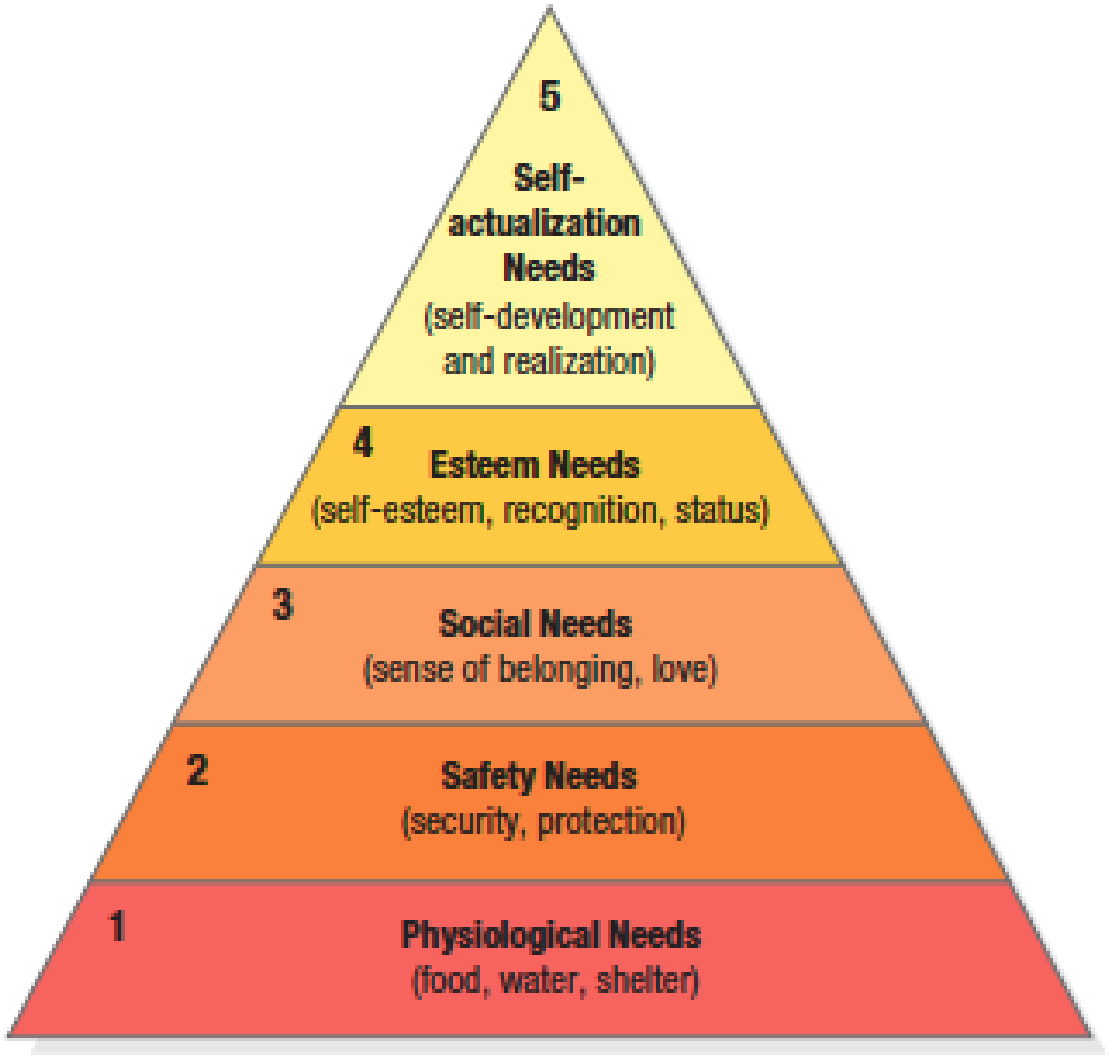
Sadness

Surprise

Fear

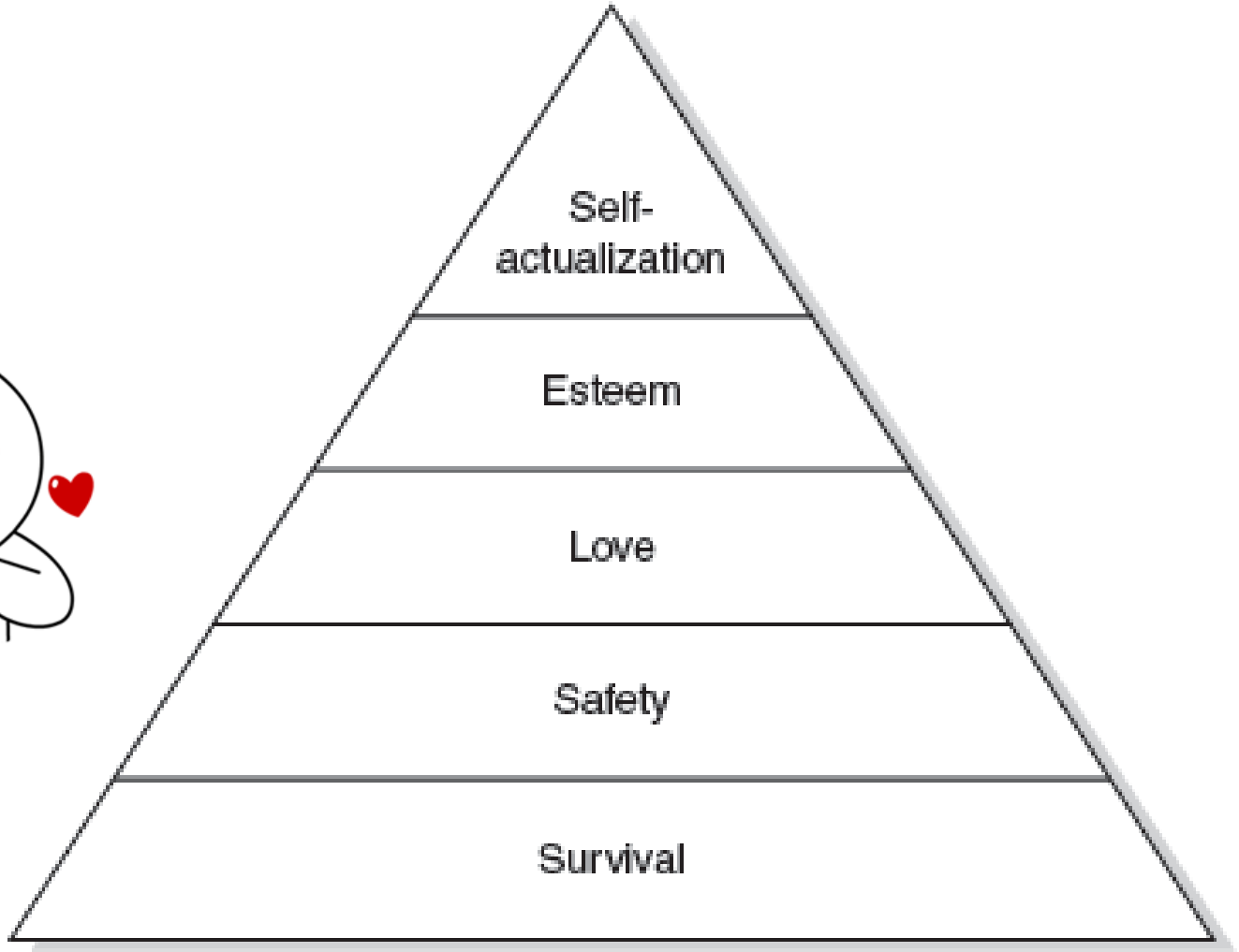
Source: Bing Liu (2011) , “Web Data Mining: Exploring Hyperlinks, Contents, and Usage Data,” Springer, 2nd Edition,

Maslow's Hierarchy of Needs



Source: Philip Kotler & Kevin Lane Keller, Marketing Management, 14th ed., Pearson, 2012

Maslow's hierarchy of human needs
(Maslow, 1943)

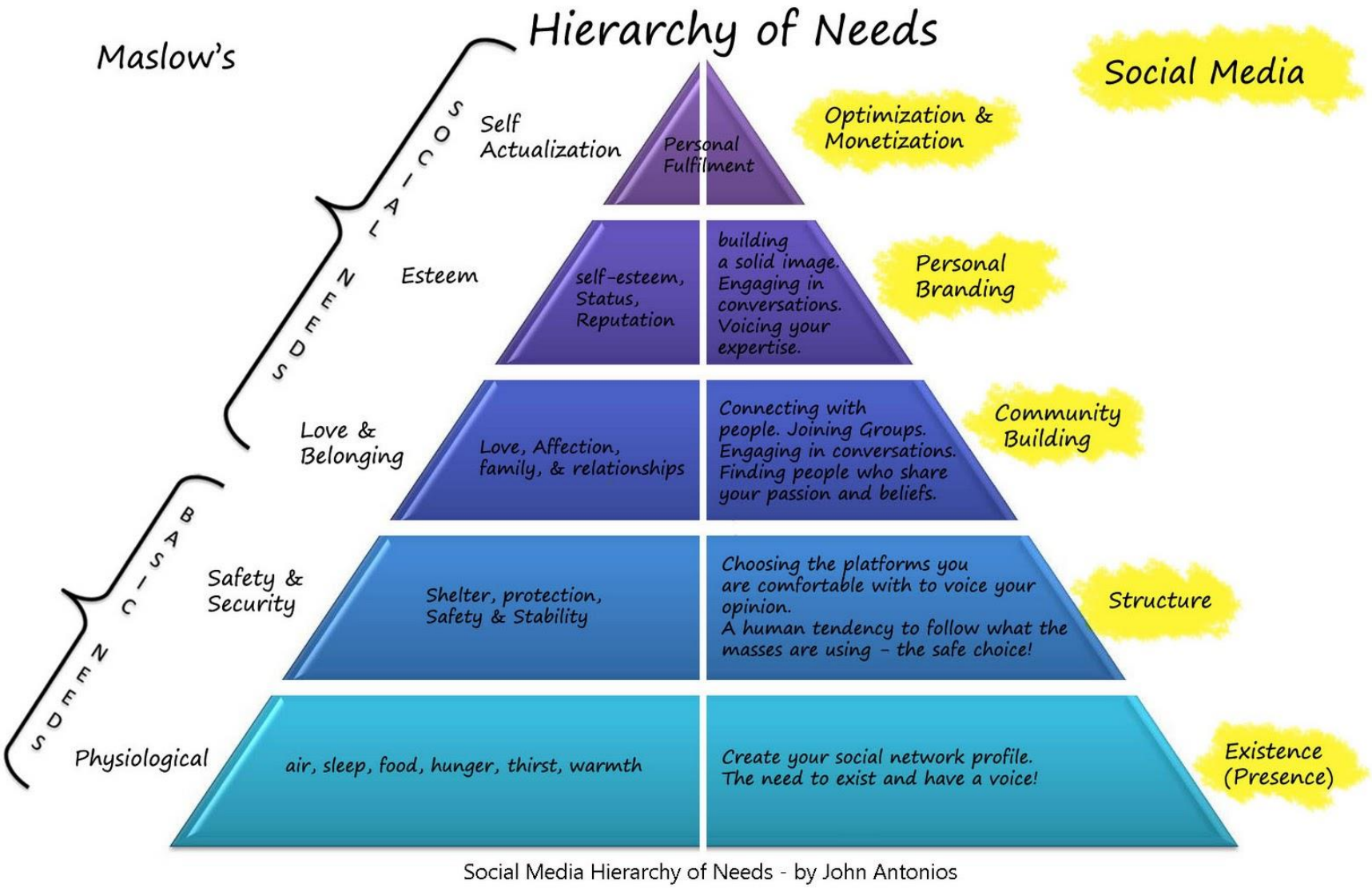


Source: Backer & Saren (2009), Marketing Theory: A Student Text, 2nd Edition, Sage

Maslow's Hierarchy of Needs

Source: http://sixstoriesup.com/social-psyche-what-makes-us-go-social/

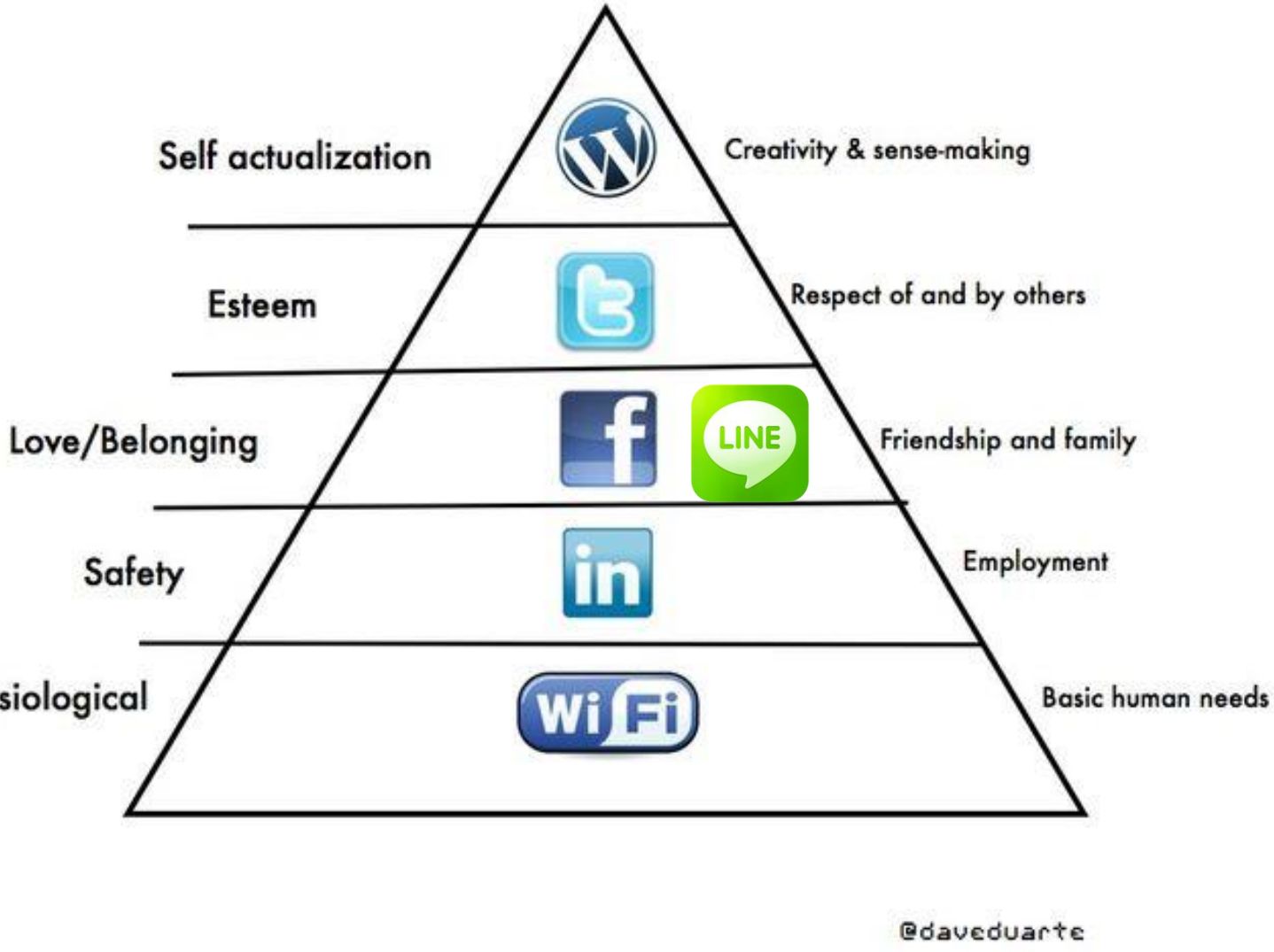
Social Media Hierarchy of Needs



Social Media Hierarchy of Needs - by John Antonios

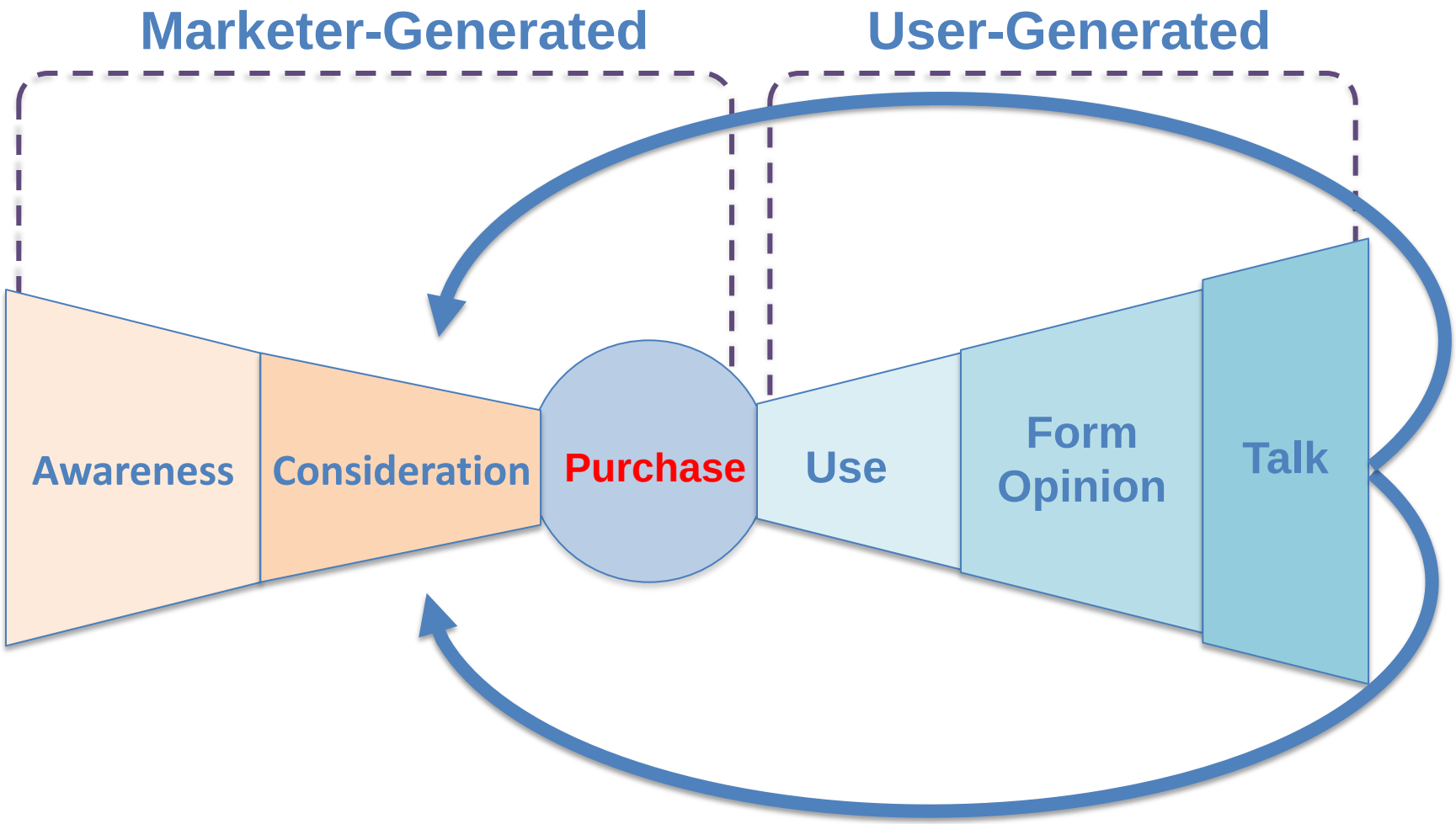
Source: http://2.bp.blogspot.com/_Rta1VZltiMk/TPavcanFtfI/AAAAAAAAACo/OBGnRL5arSU/s1600/social-media-heirarchy-of-needs1.jpg

Social Media Hierarchy of Needs



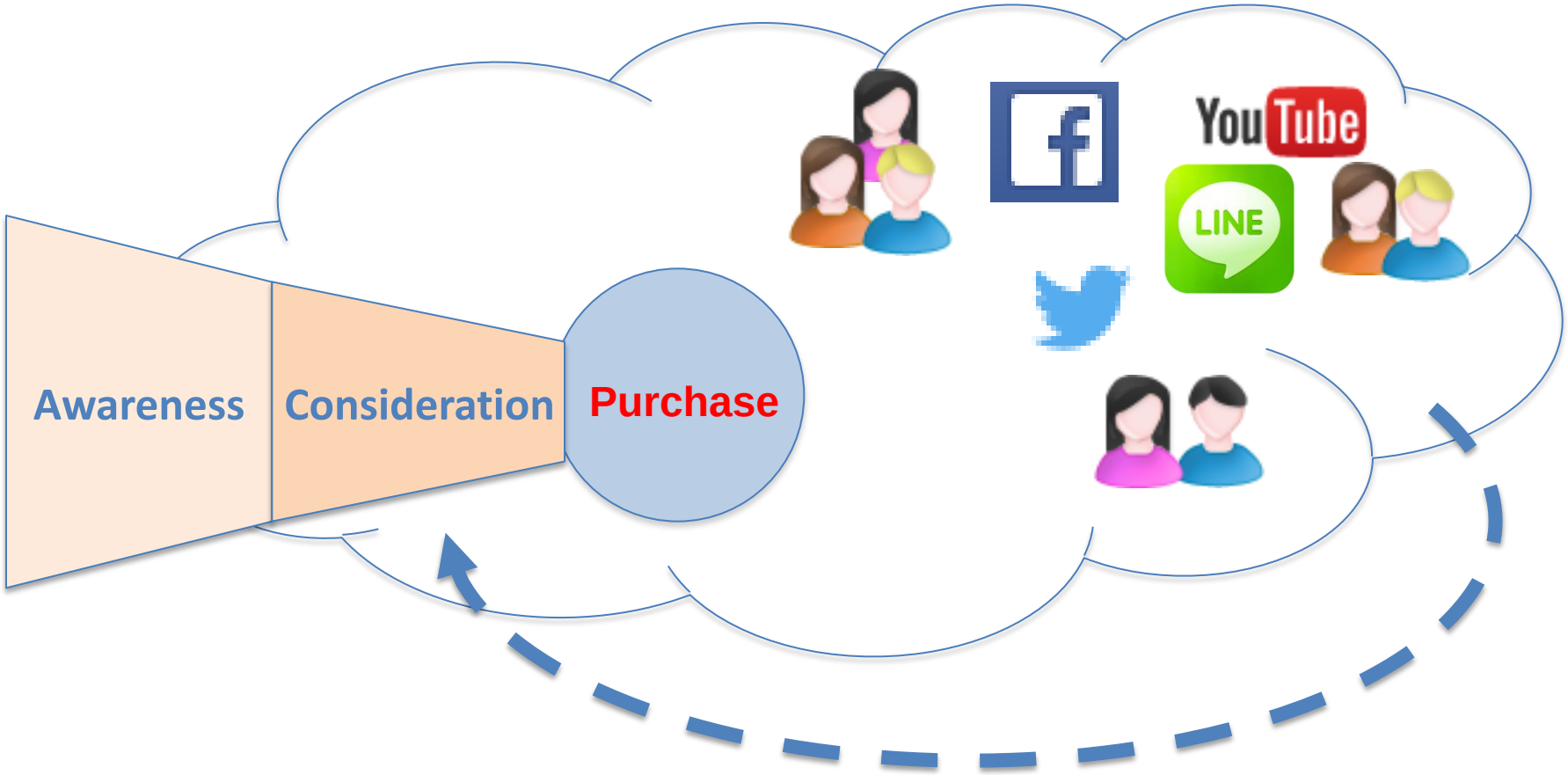
Source: http://www.pinterest.com/pin/18647785930903585/

The Social Feedback Cycle
Consumer Behavior on Social Media



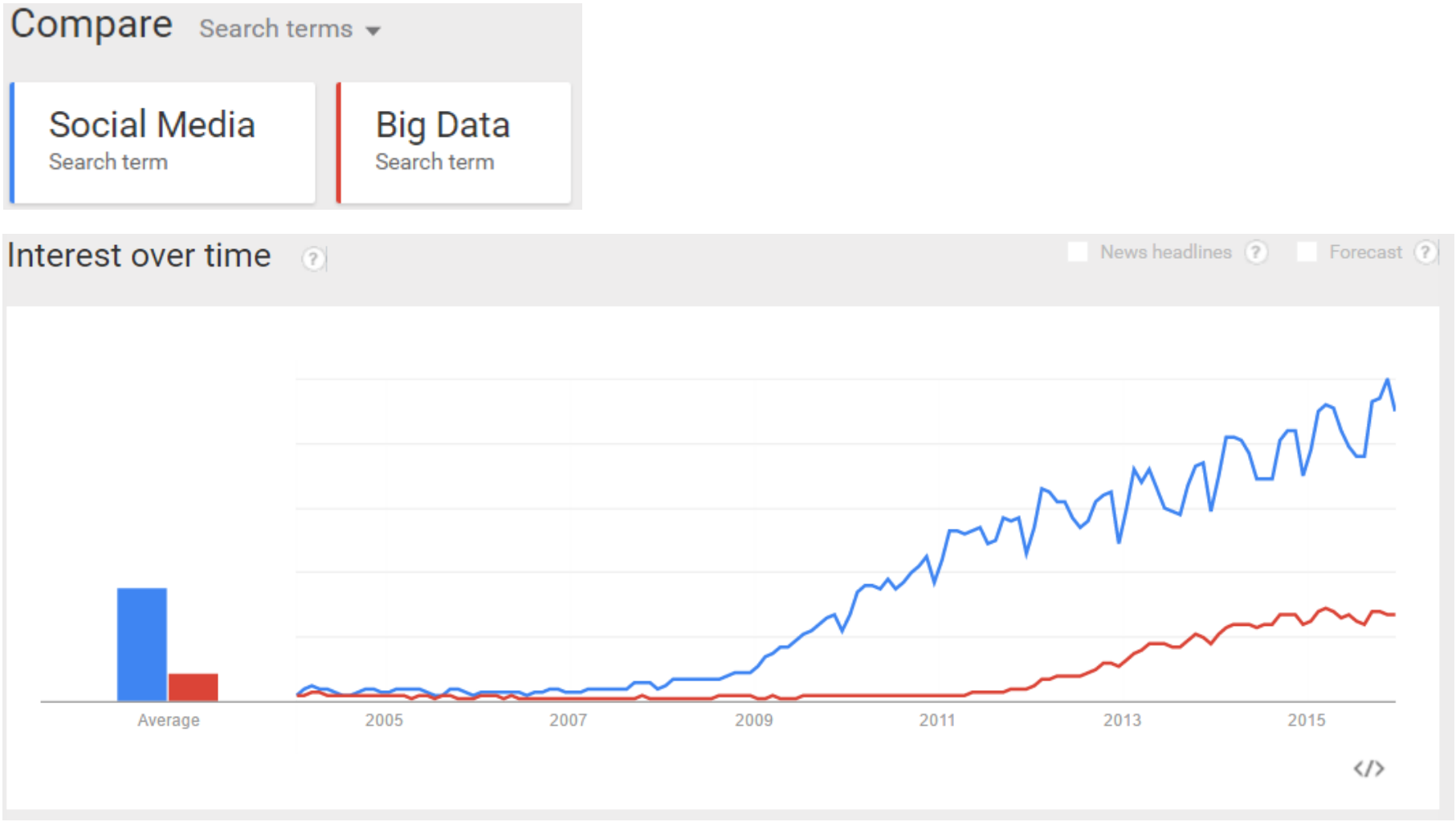
Source: Evans et al. (2010), Social Media Marketing: The Next Generation of Business Engagement

The New Customer Influence Path



Source: Evans et al. (2010), Social Media Marketing: The Next Generation of Business Engagement

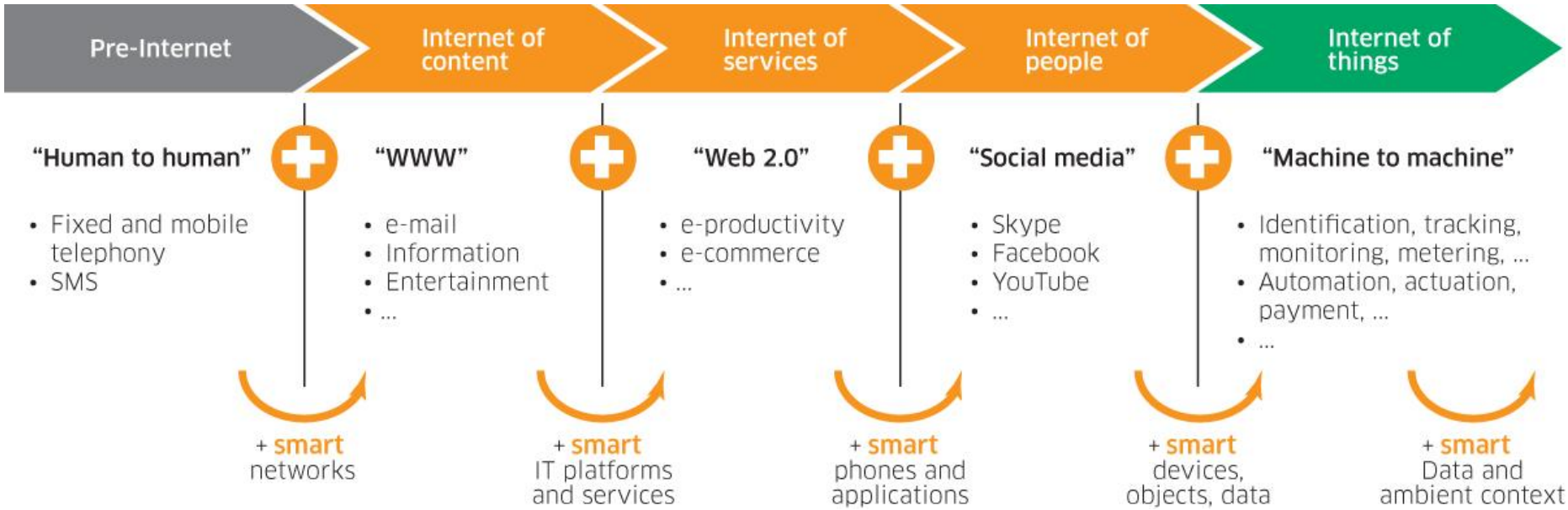
Google Trends on Social Media



Source: http://www.google.com.tw/trends/explore#q=Social%20Media%2C%20Big%20Data

Internet Evolution

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



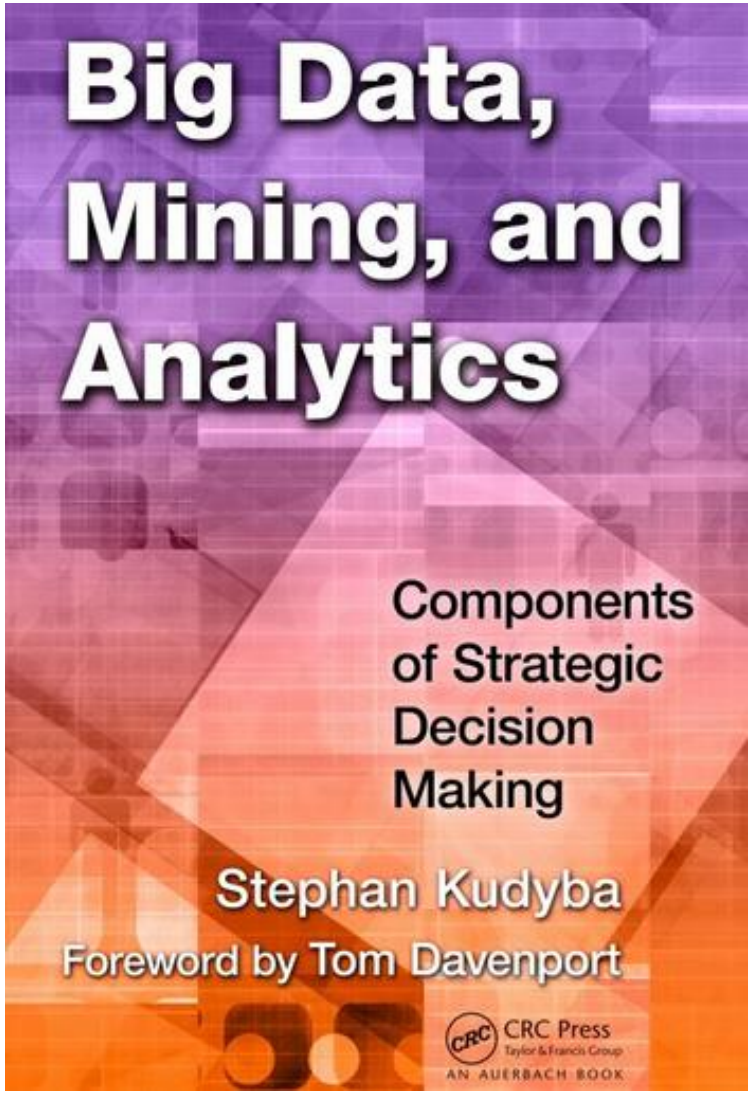
Source: Marc Jadoul (2015), The IoT: The next step in internet evolution, March 11, 2015
http://www2.alcatel-lucent.com/techzine/iot-internet-of-things-next-step-evolution/

Business Insights with Social Analytics

Big Data Analytics and Data Mining

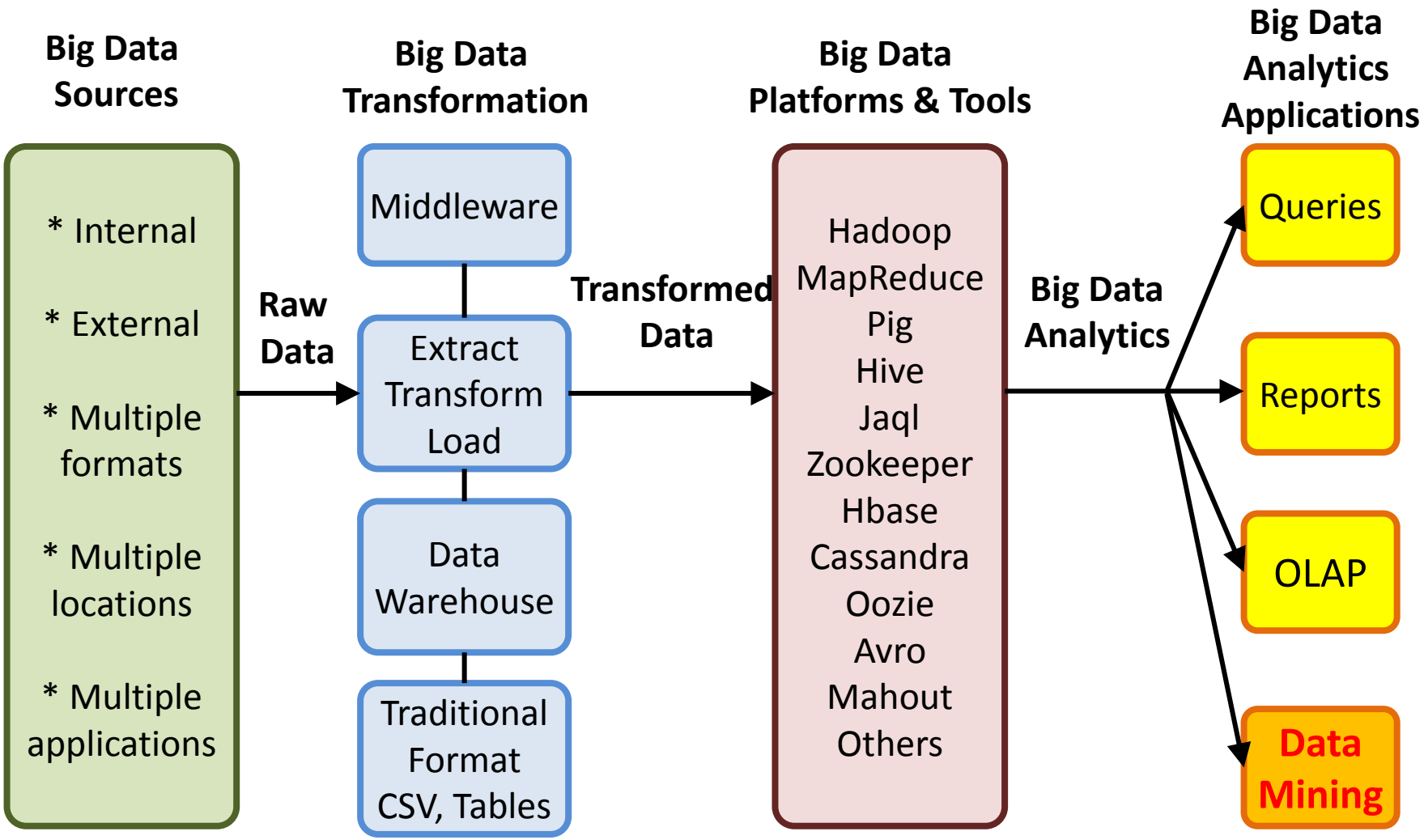
Stephan Kudyba (2014),

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



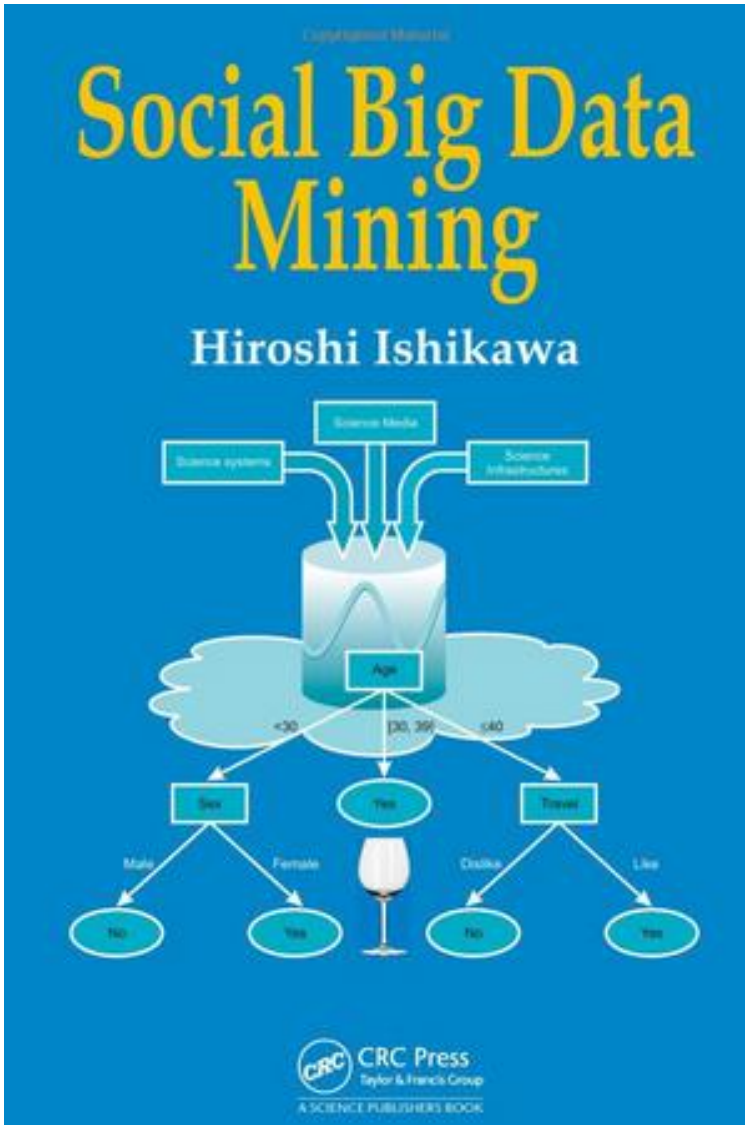
Source: Stephan Kudyba (2014), Big Data, Mining, and Analytics: Components of Strategic Decision Making, Auerbach Publications

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)



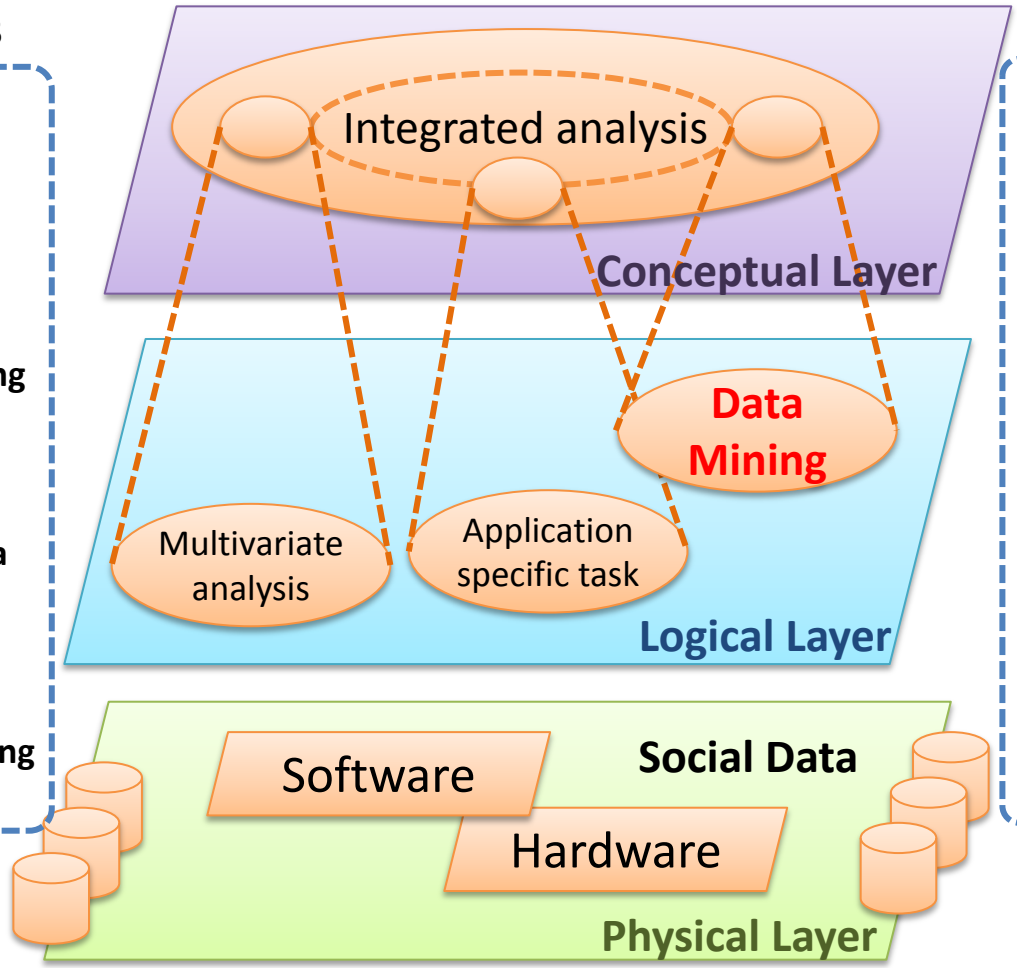
Source: http://www.amazon.com/Social-Data-Mining-Hiroshi-Ishikawa/dp/149871093X

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

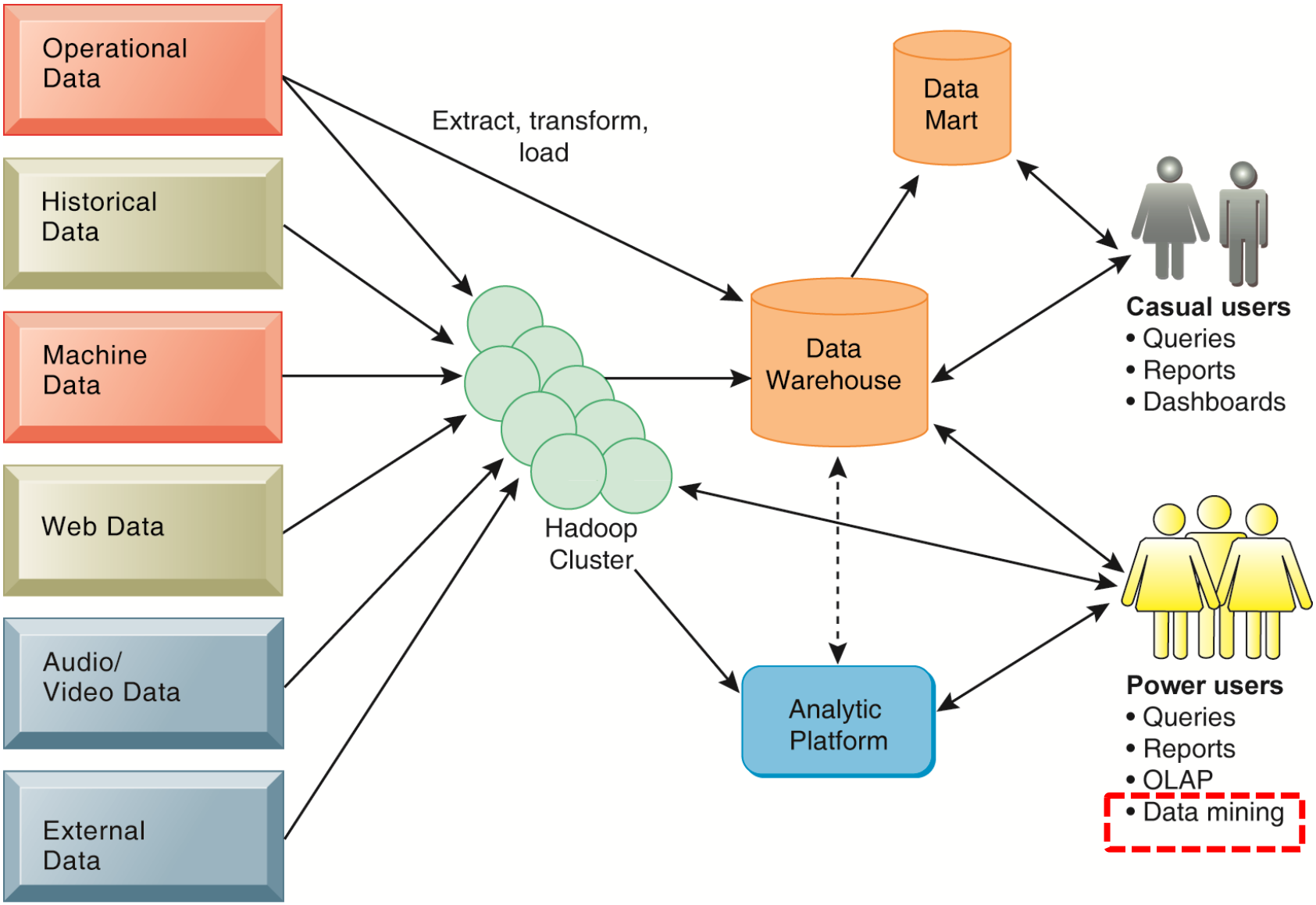


Analysts

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

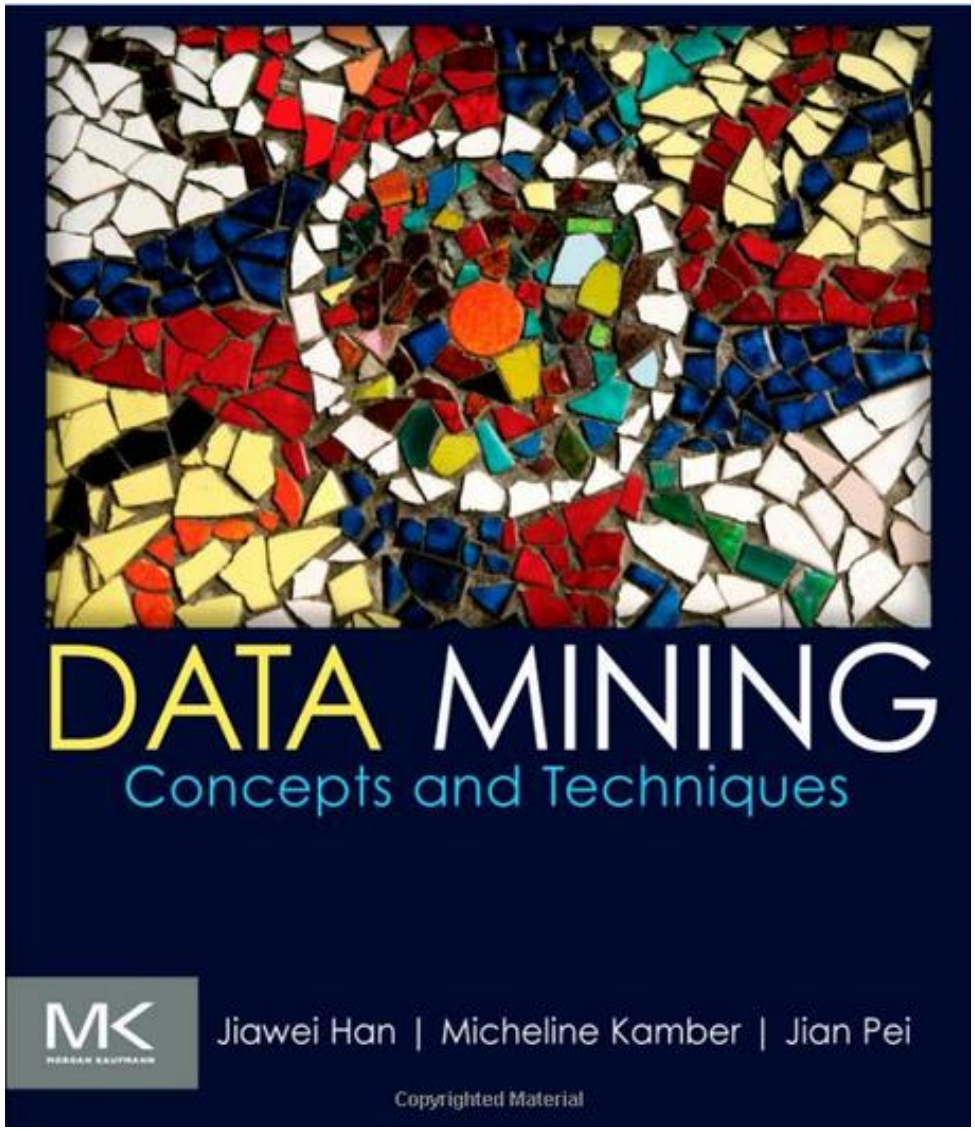
Source: Hiroshi Ishikawa (2015), Social Big Data Mining, CRC Press

Business Intelligence (BI) Infrastructure



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

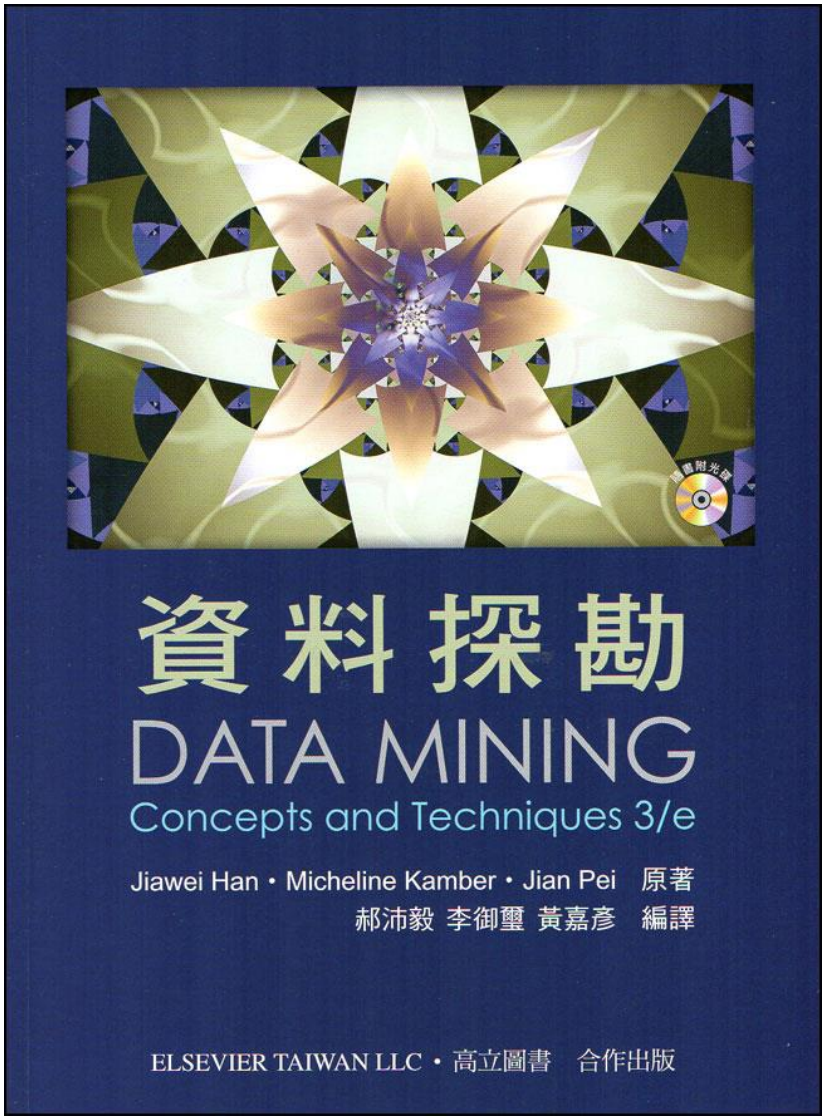
Data Mining



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

郝沛毅, 李御璽, 黃嘉彥 編譯, 資料探勘

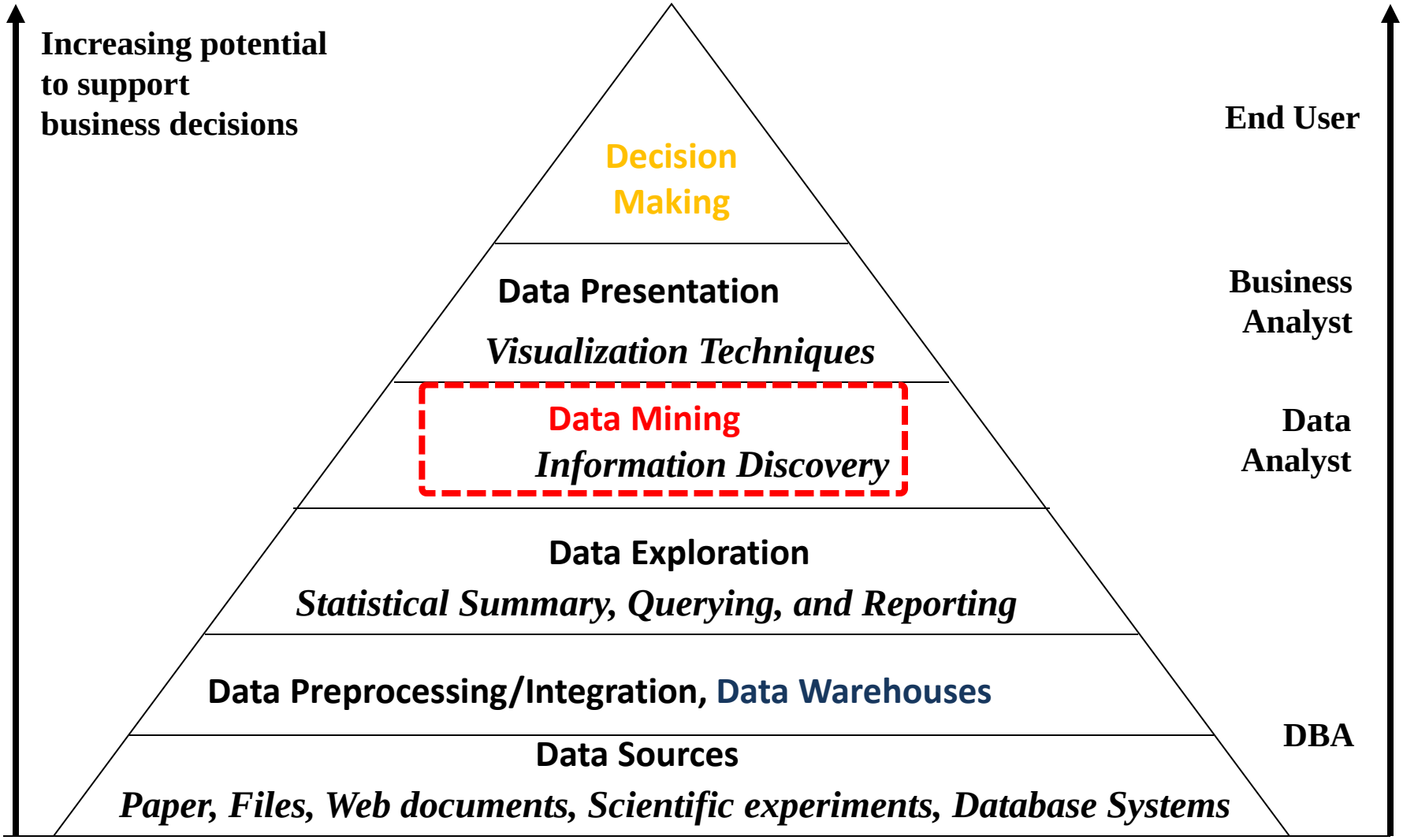
(Jiawei Han, Micheline Kamber, Jian Pei, Data Mining - Concepts and Techniques 3/e), 高立圖書, 2014



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

Data Warehouse

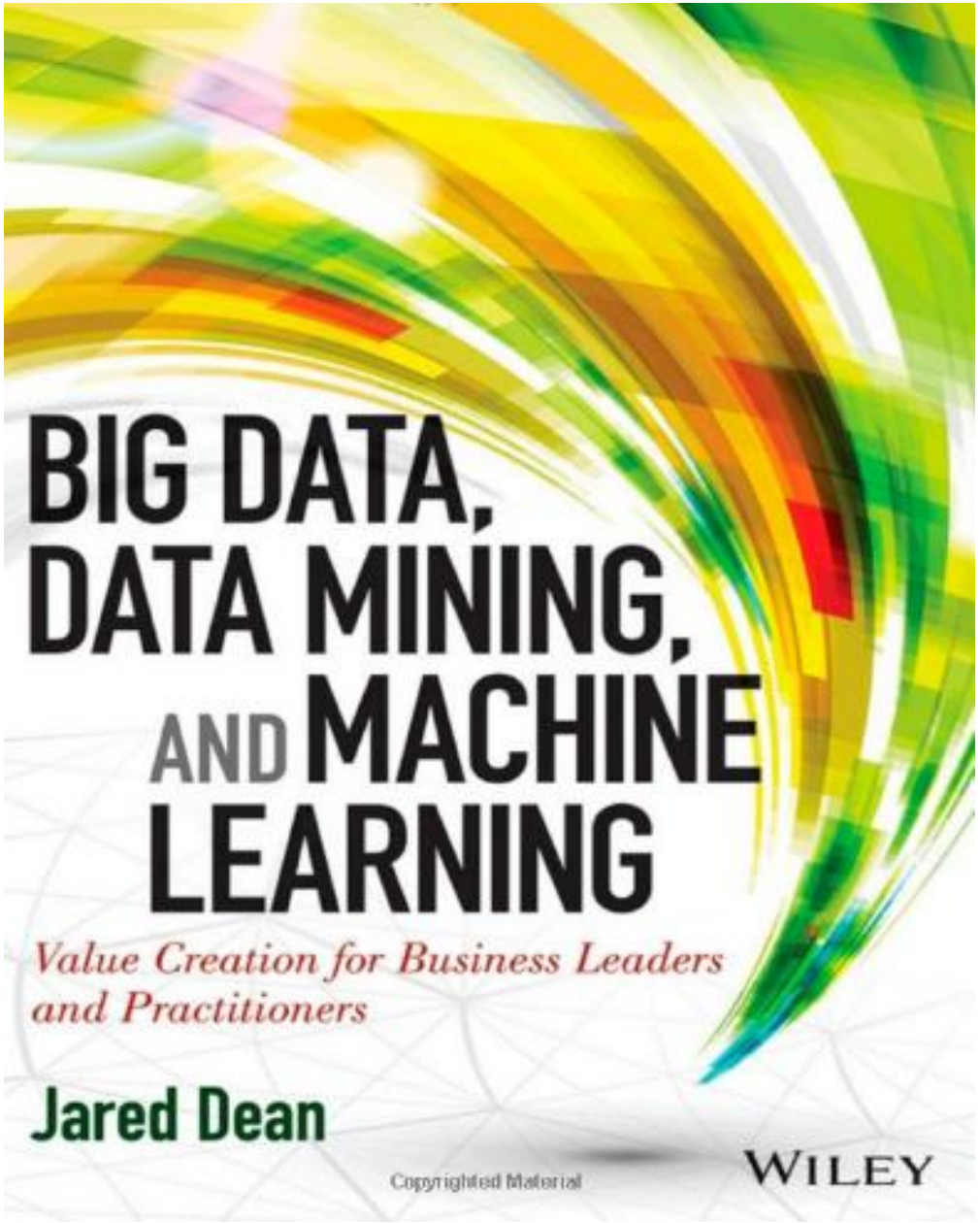
Data Mining and Business Intelligence



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

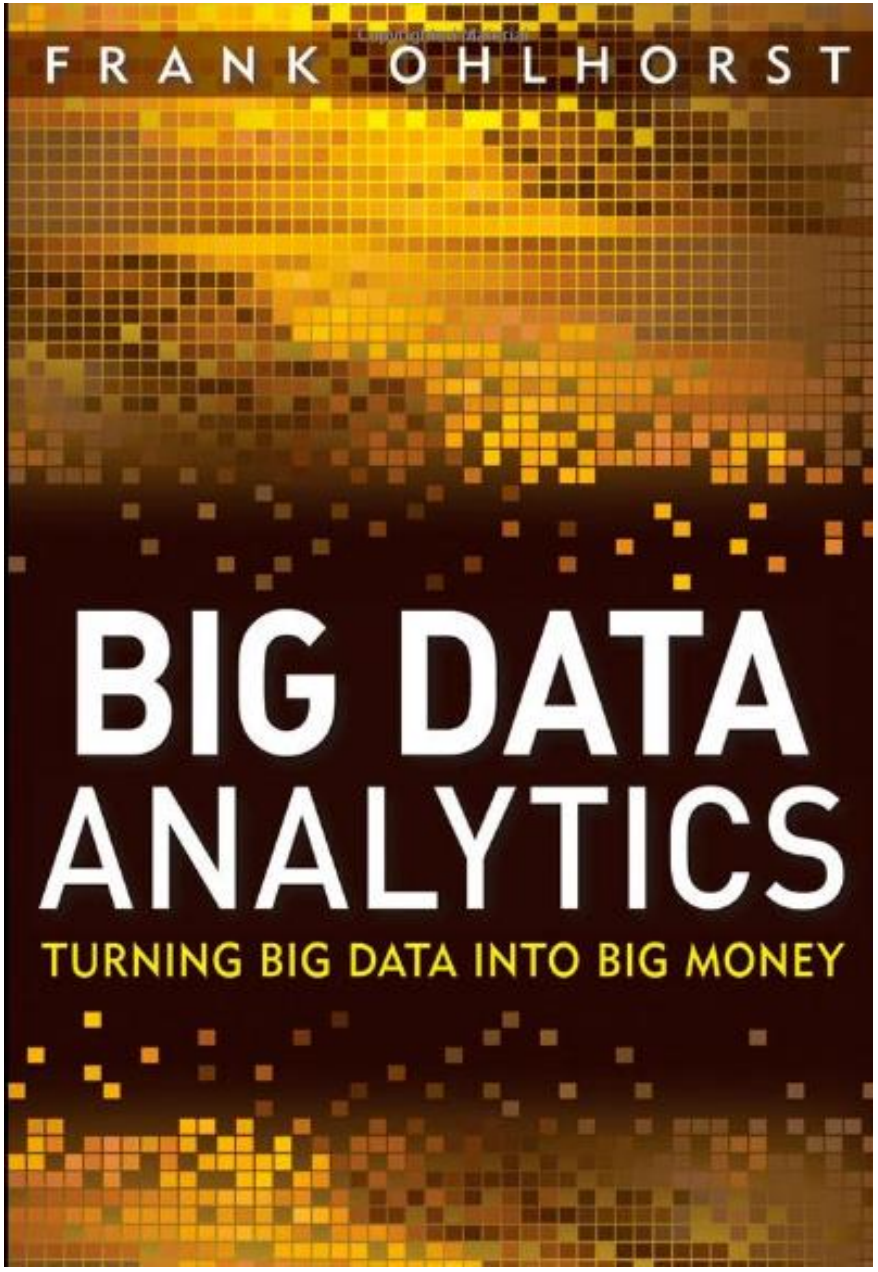


Source: http://www.amazon.com/Data-Mining-Machine-Learning-Practitioners/dp/1118618041

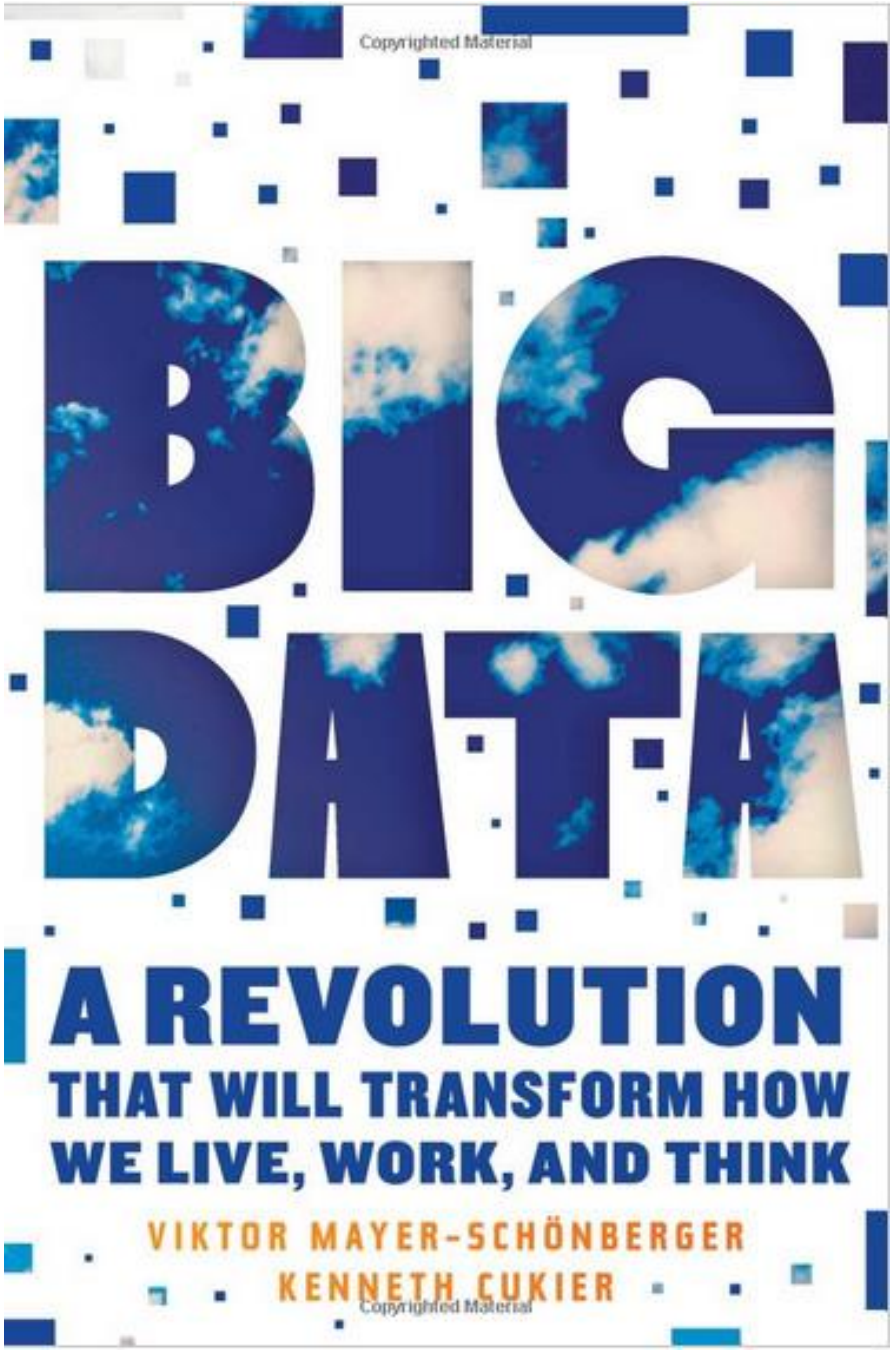
Deep Learning

Intelligence from Big Data

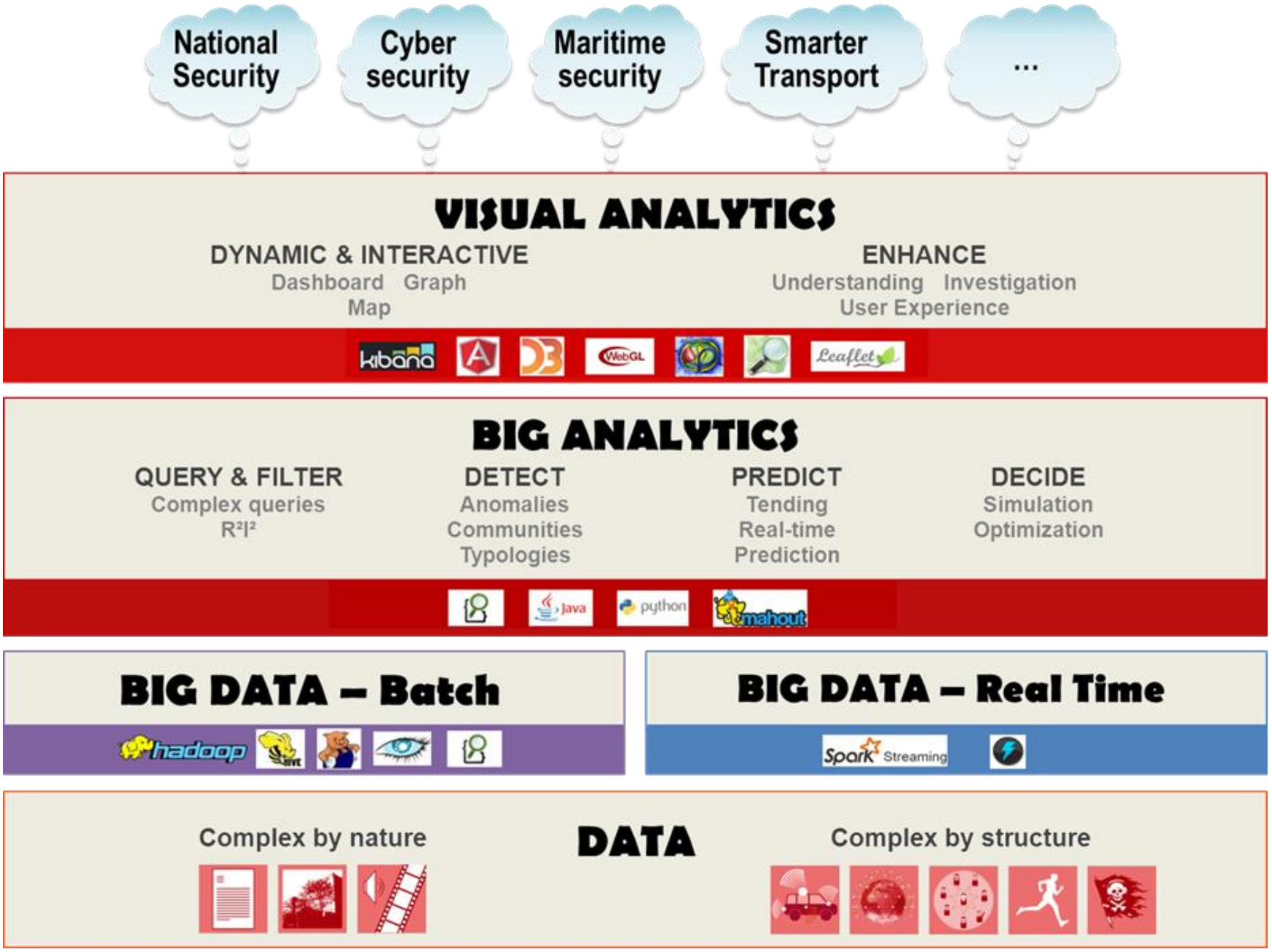
Source: https://www.vlab.org/events/deep-learning/



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

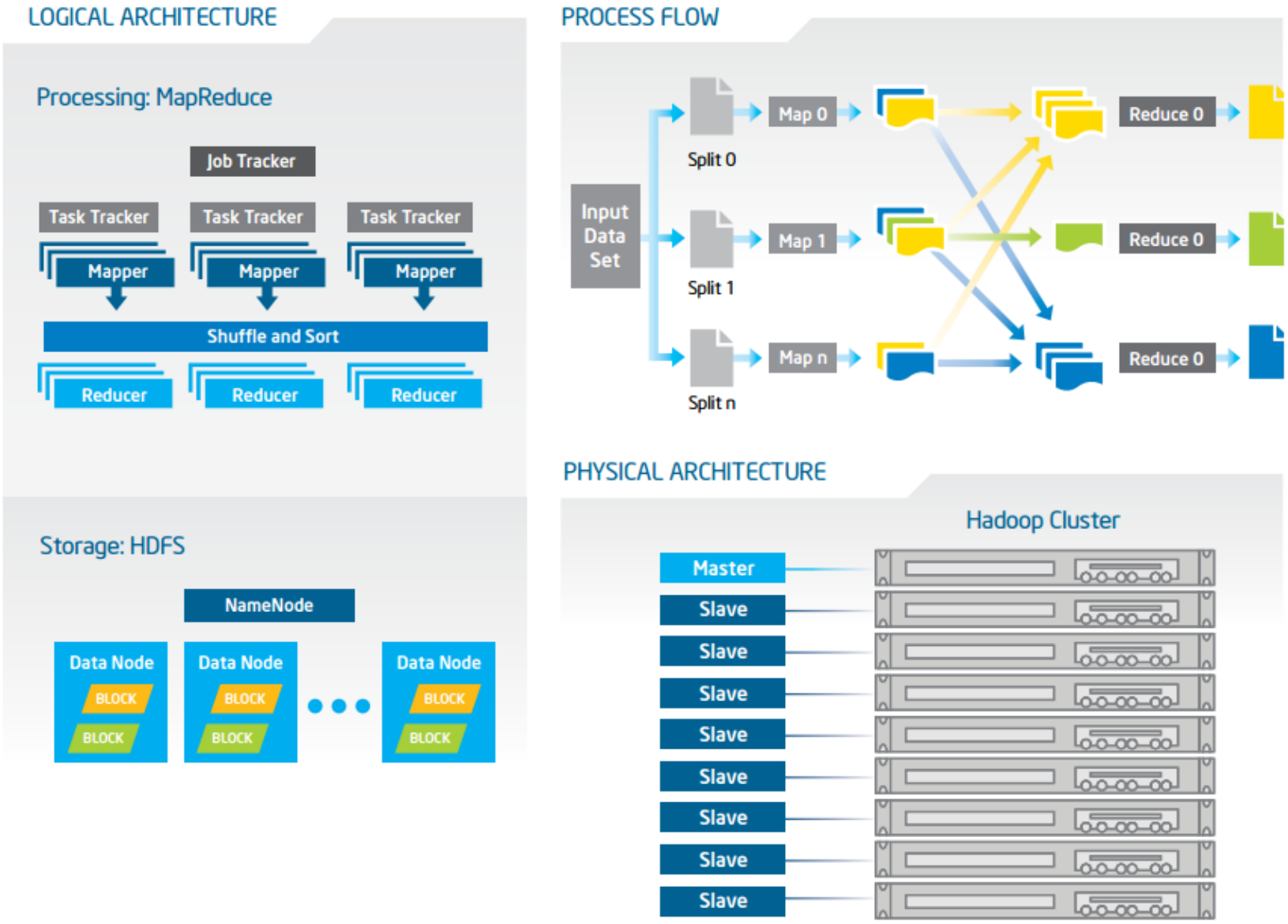


Source: http://www.amazon.com/Big-Data-Revolution-Transform-Mayer-Schonberger/dp/B00D81X2YE



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

Big Data with Hadoop Architecture

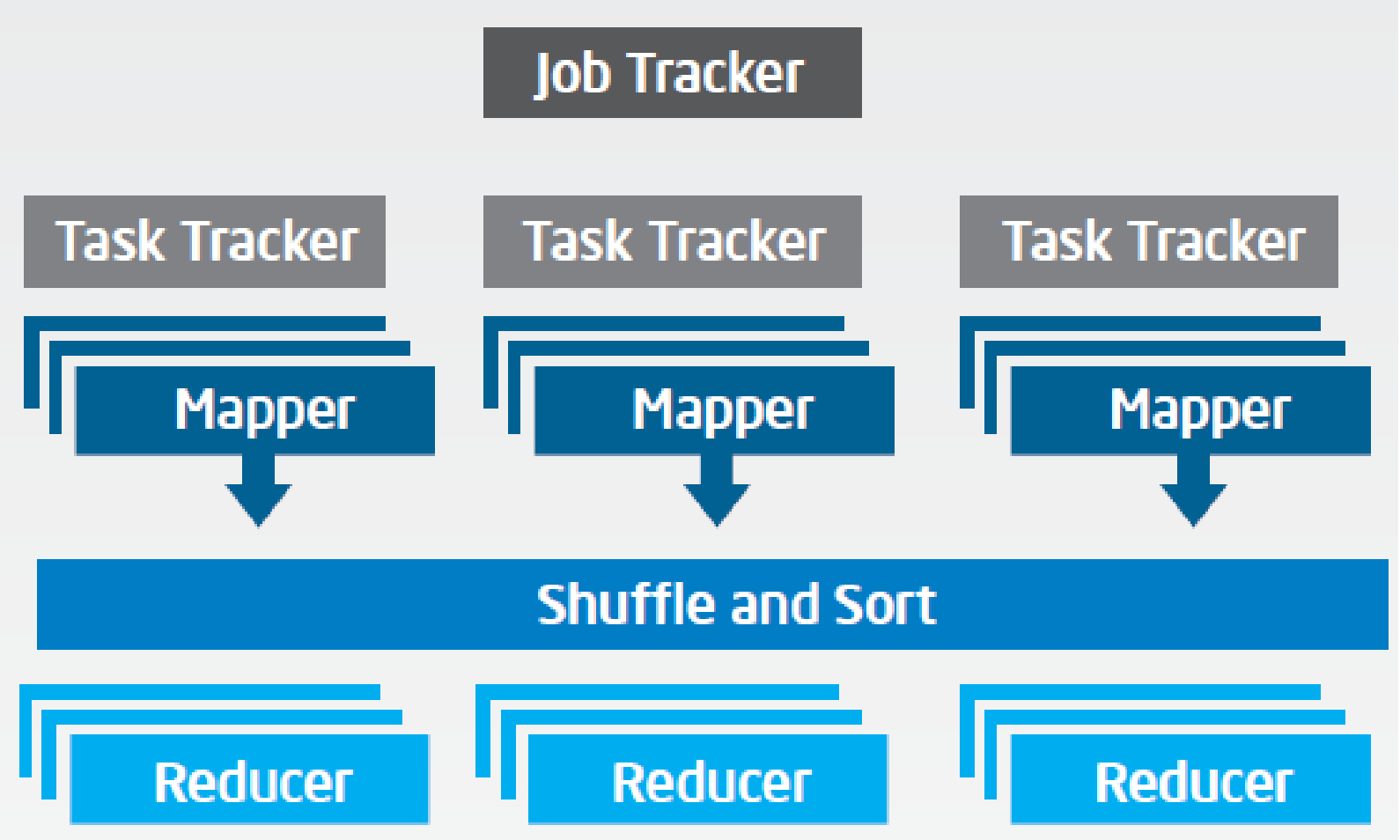


Source: https://software.intel.com/sites/default/files/article/402274/etl-big-data-with-hadoop.pdf

Big Data with Hadoop Architecture

Logical Architecture

Processing: MapReduce

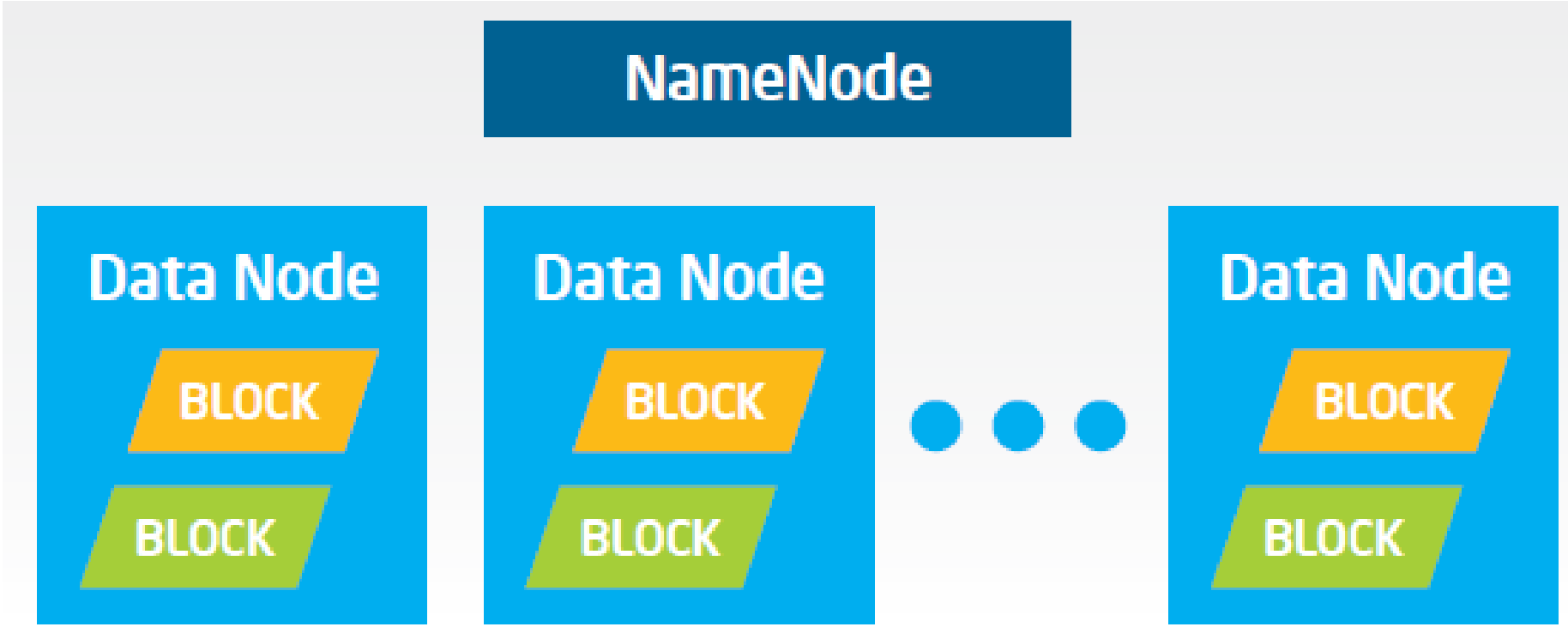


Source: https://software.intel.com/sites/default/files/article/402274/etl-big-data-with-hadoop.pdf

Big Data with Hadoop Architecture

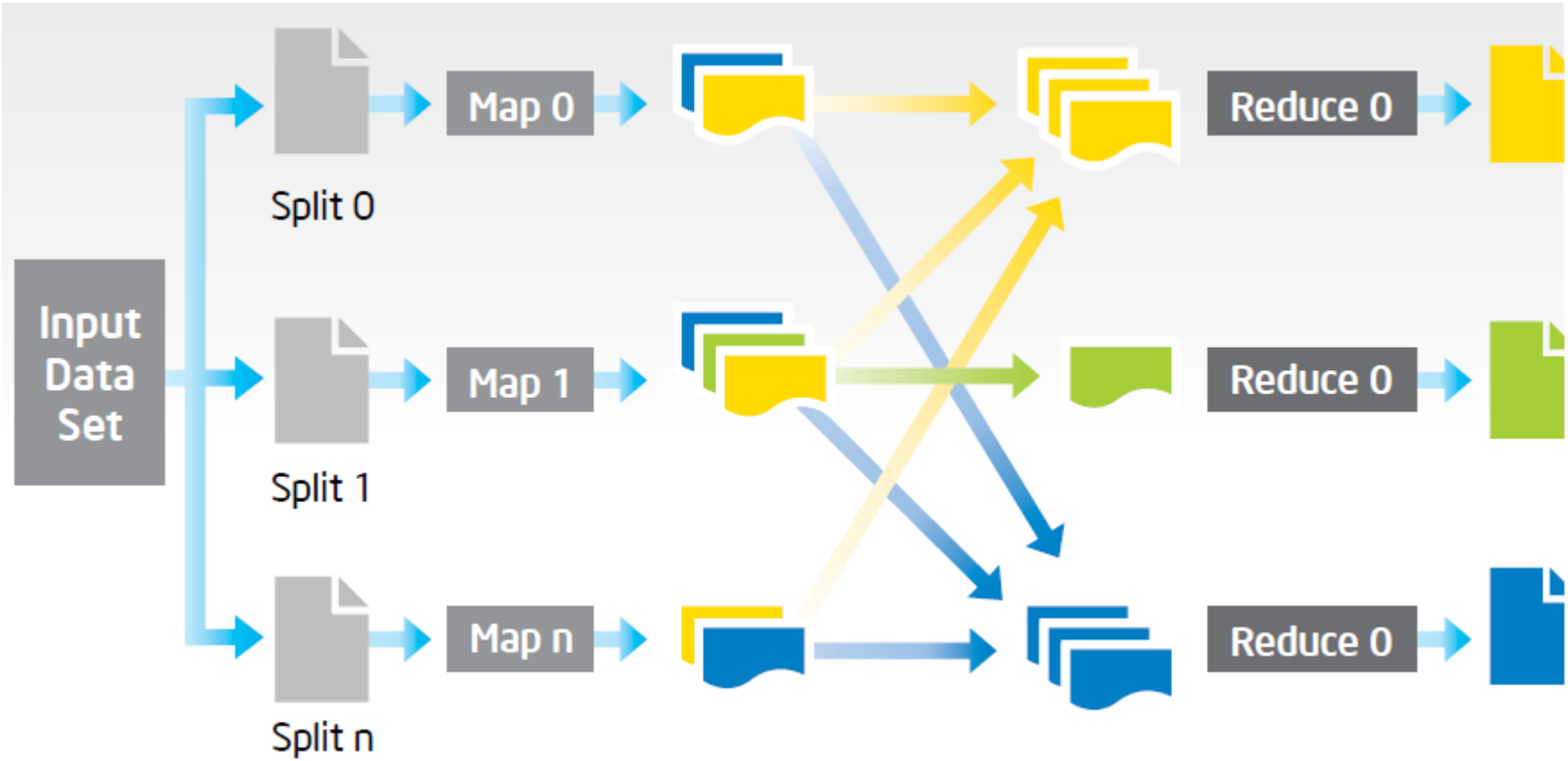
Logical Architecture

Storage: HDFS



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

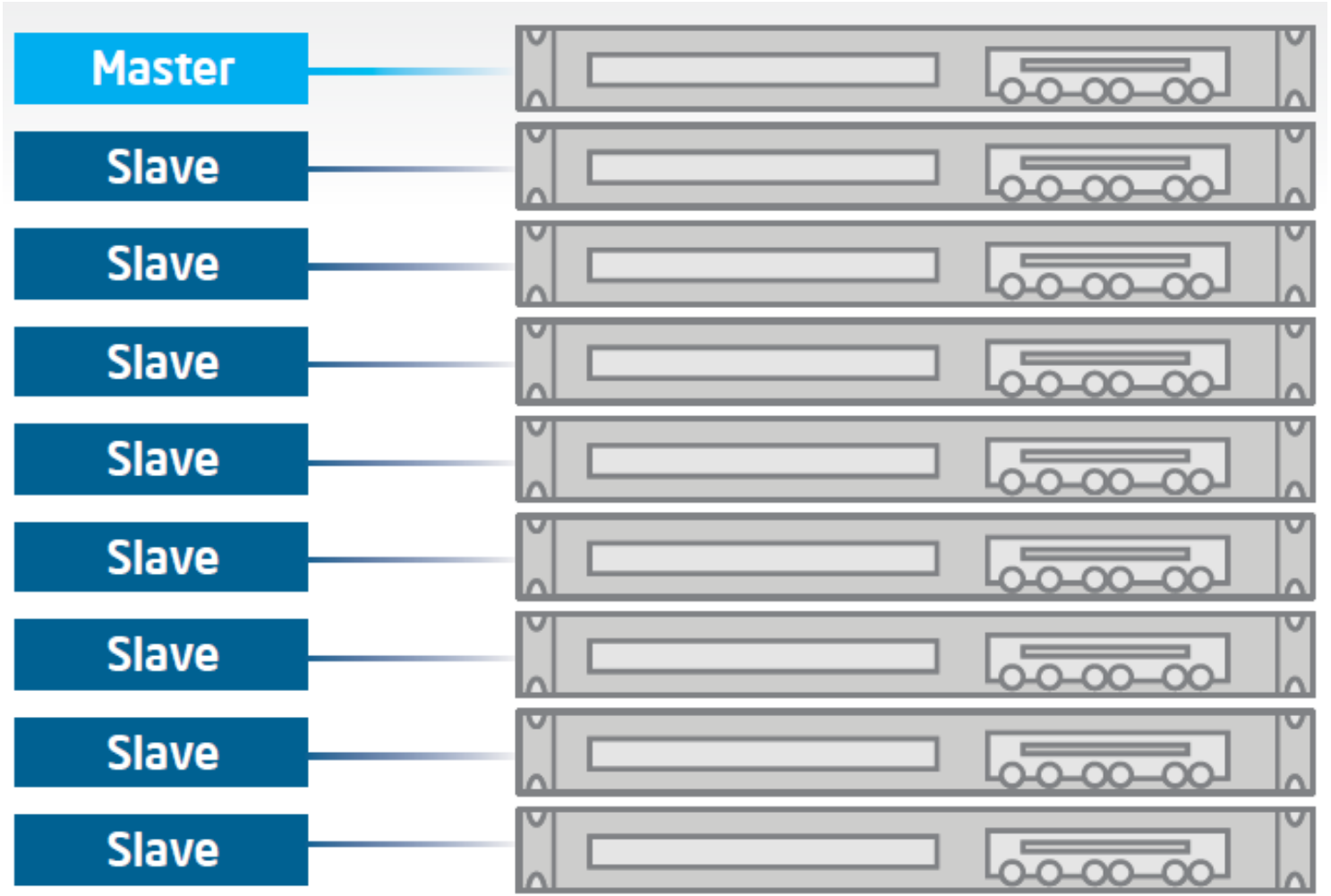
Big Data with Hadoop Architecture
Process Flow



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

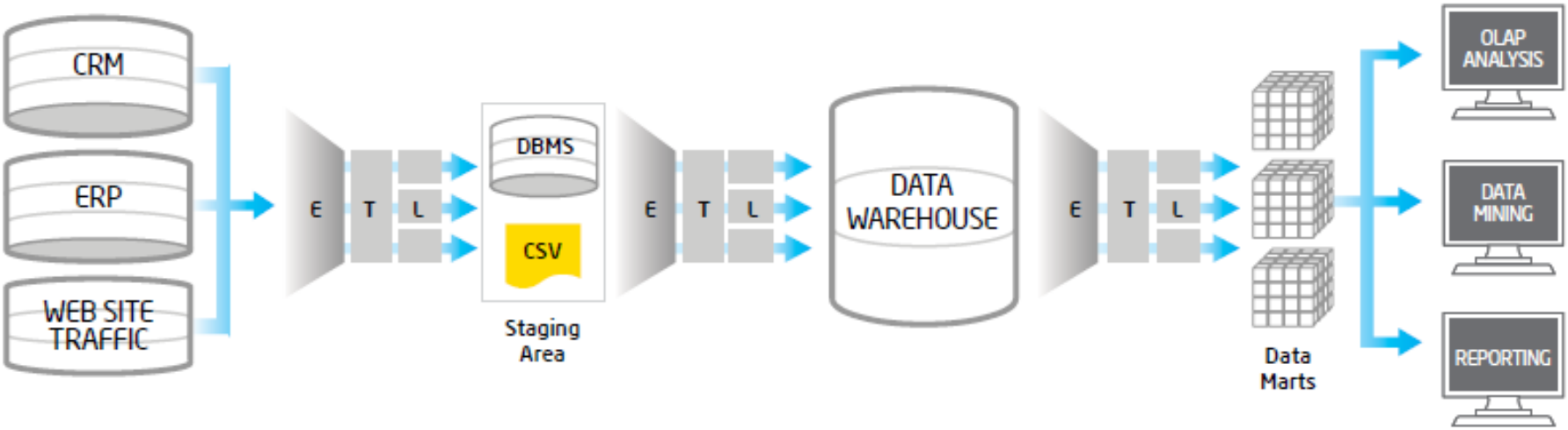
Big Data with Hadoop Architecture

Hadoop Cluster



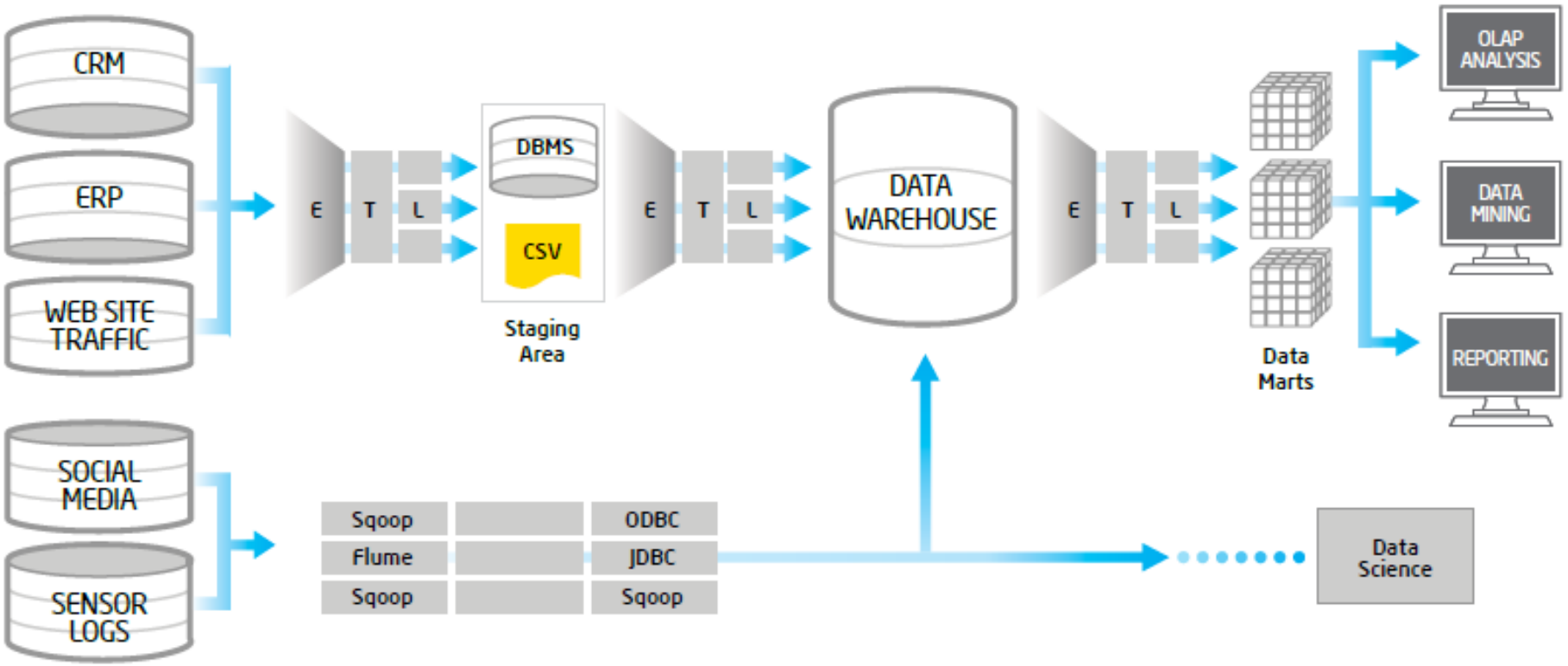
Source: https://software.intel.com/sites/default/files/article/402274/etl-big-data-with-hadoop.pdf

Traditional ETL Architecture



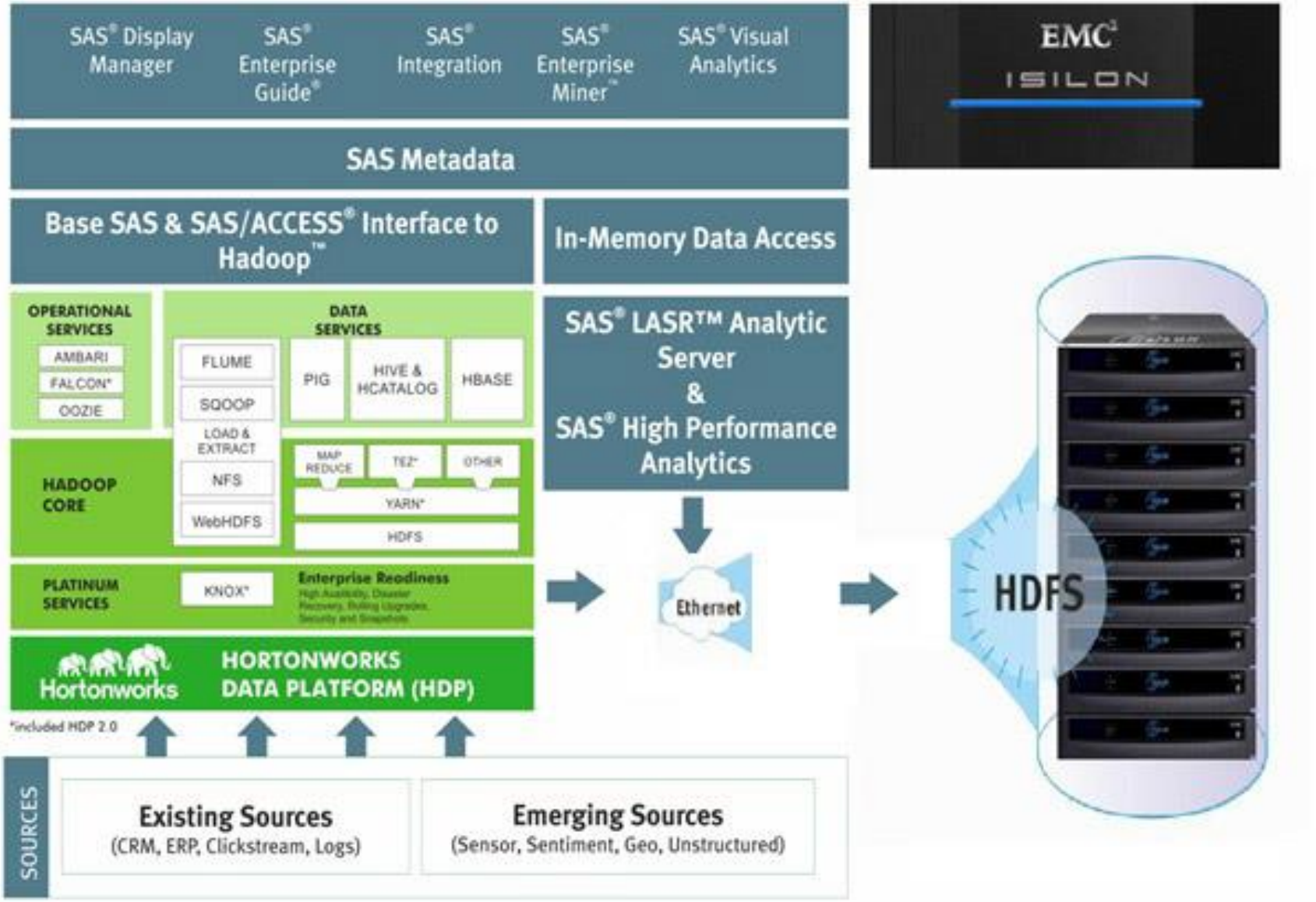
Source: https://software.intel.com/sites/default/files/article/402274/etl-big-data-with-hadoop.pdf

Offload ETL with Hadoop (Big Data Architecture)



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

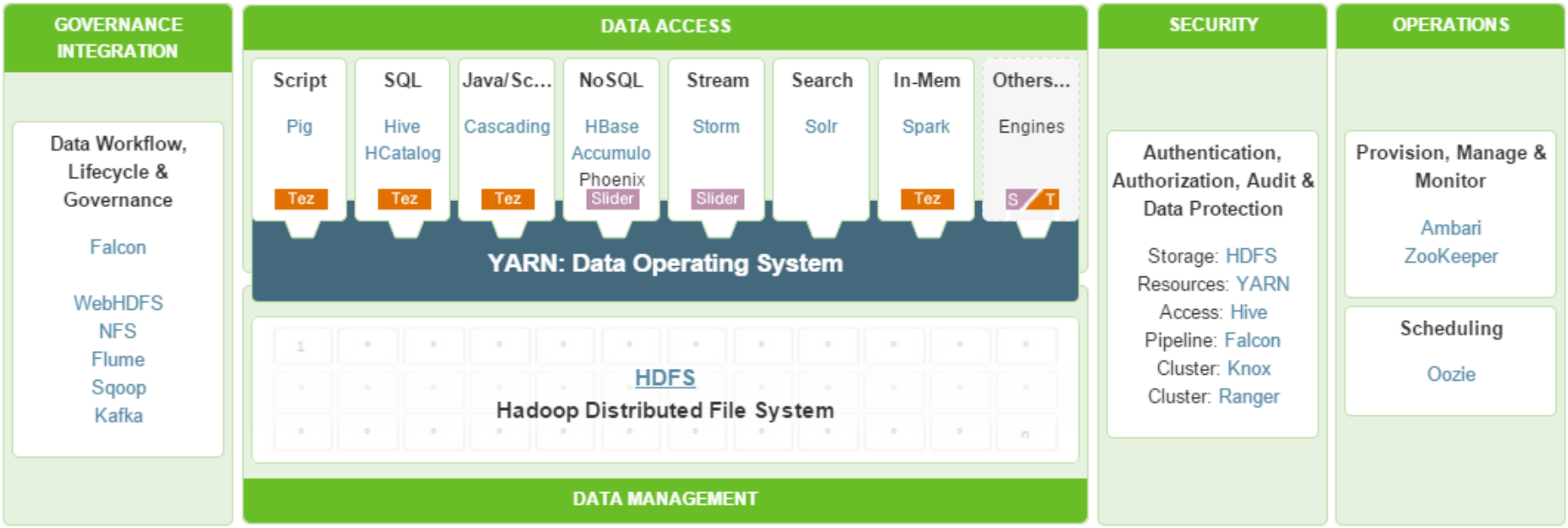
Big Data Solution



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

HDP

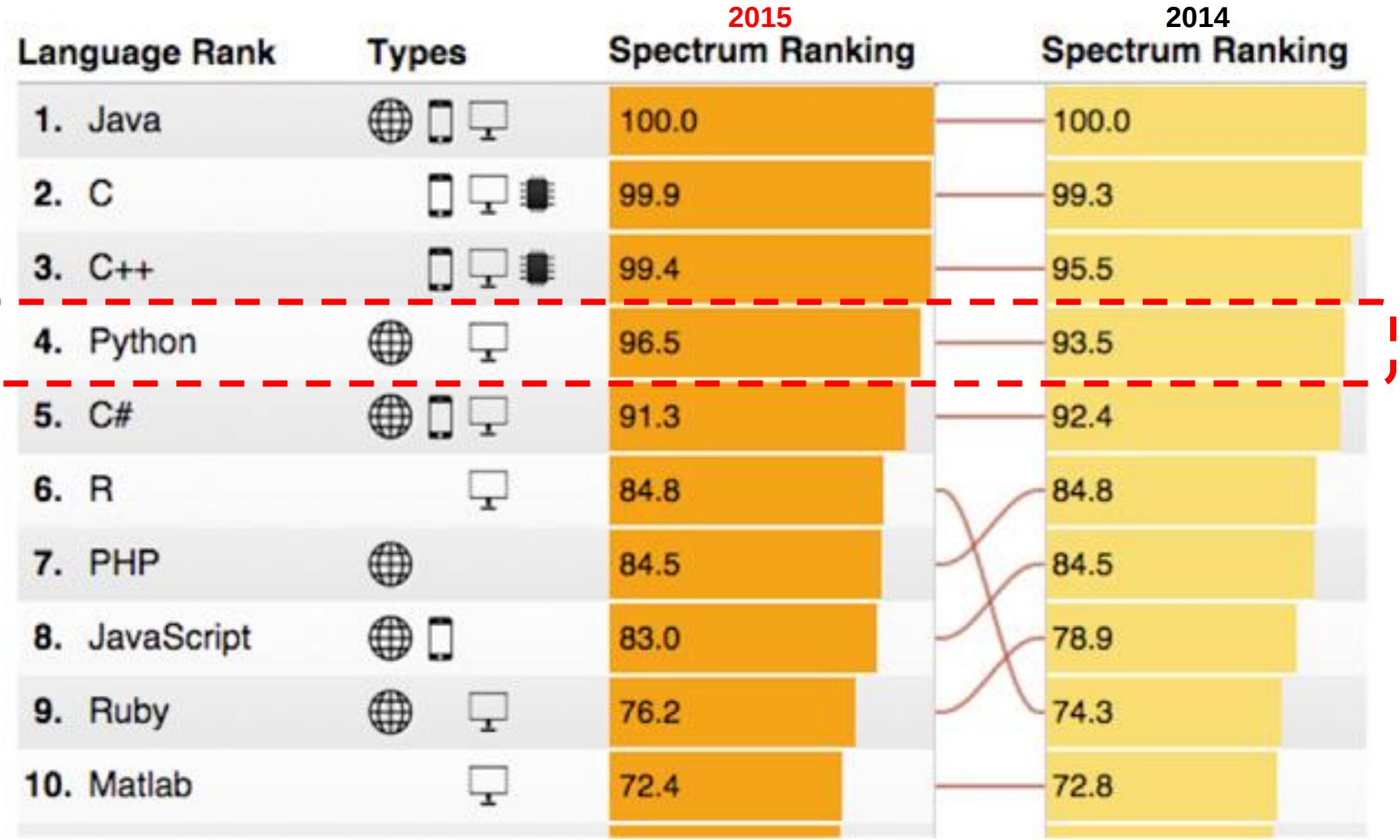
A Complete Enterprise Hadoop Data Platform



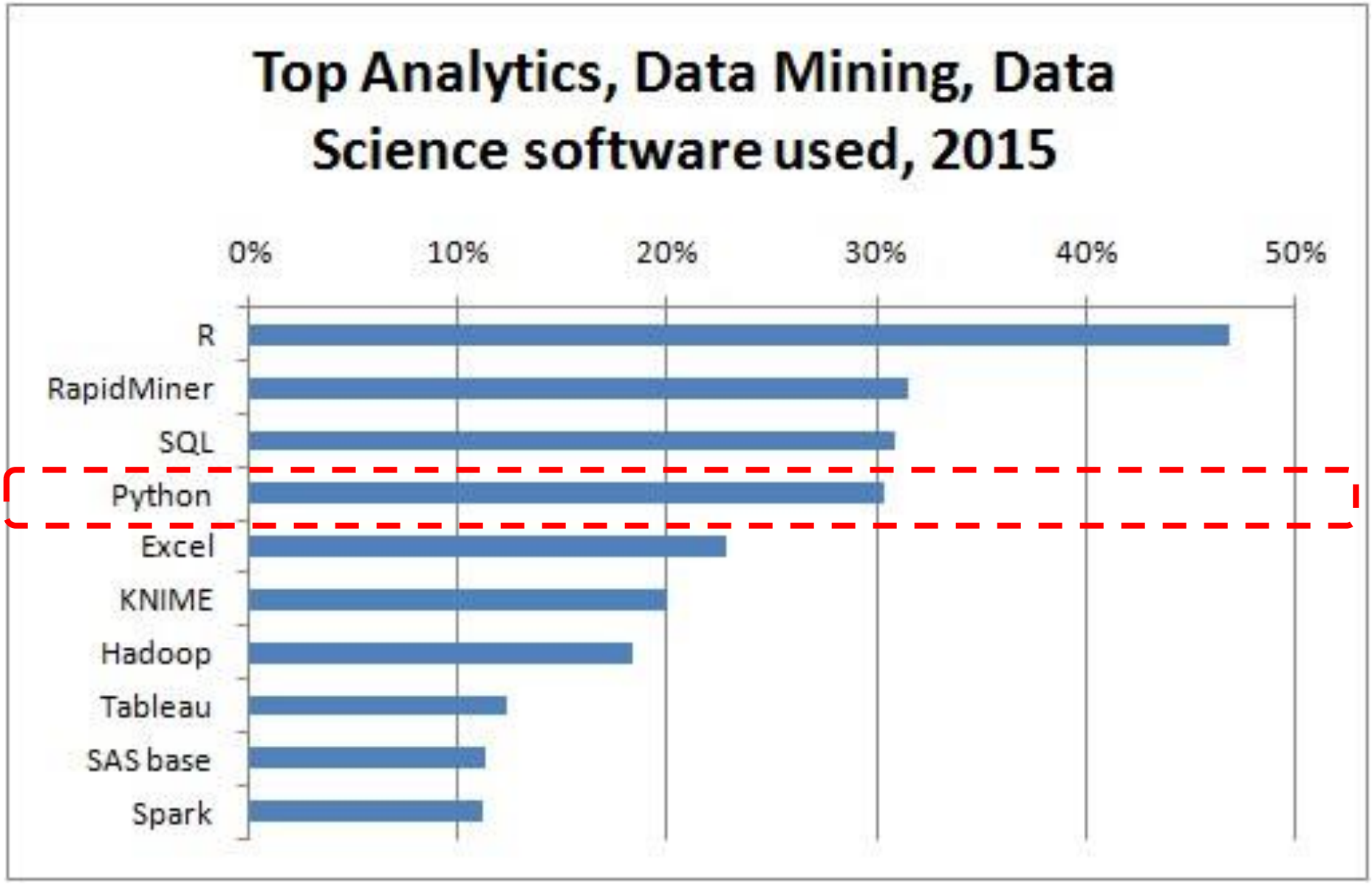
Source: http://hortonworks.com/hdp/

Python for Big Data Analytics

(The column on the left is the 2015 ranking; the column on the right is the 2014 ranking for comparison

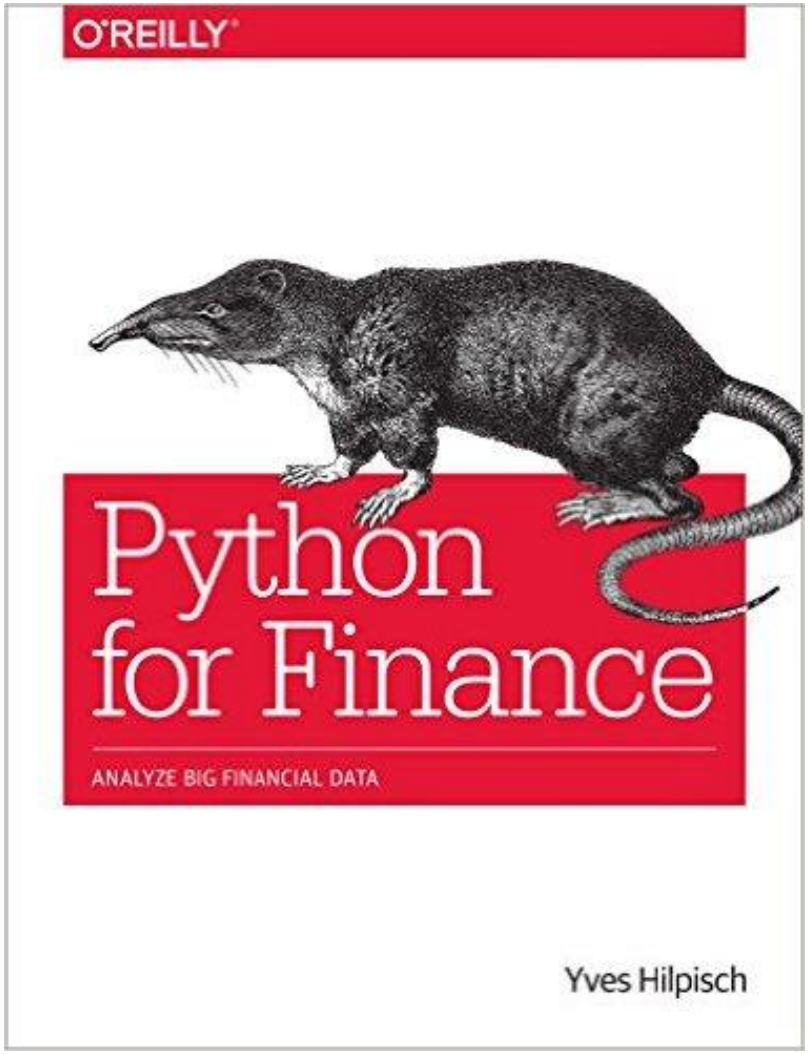


Source: http://spectrum.ieee.org/computing/software/the-2015-top-ten-programming-languages



Source: http://www.kdnuggets.com/2015/05/poll-r-rapidminer-python-big-data-spark.html

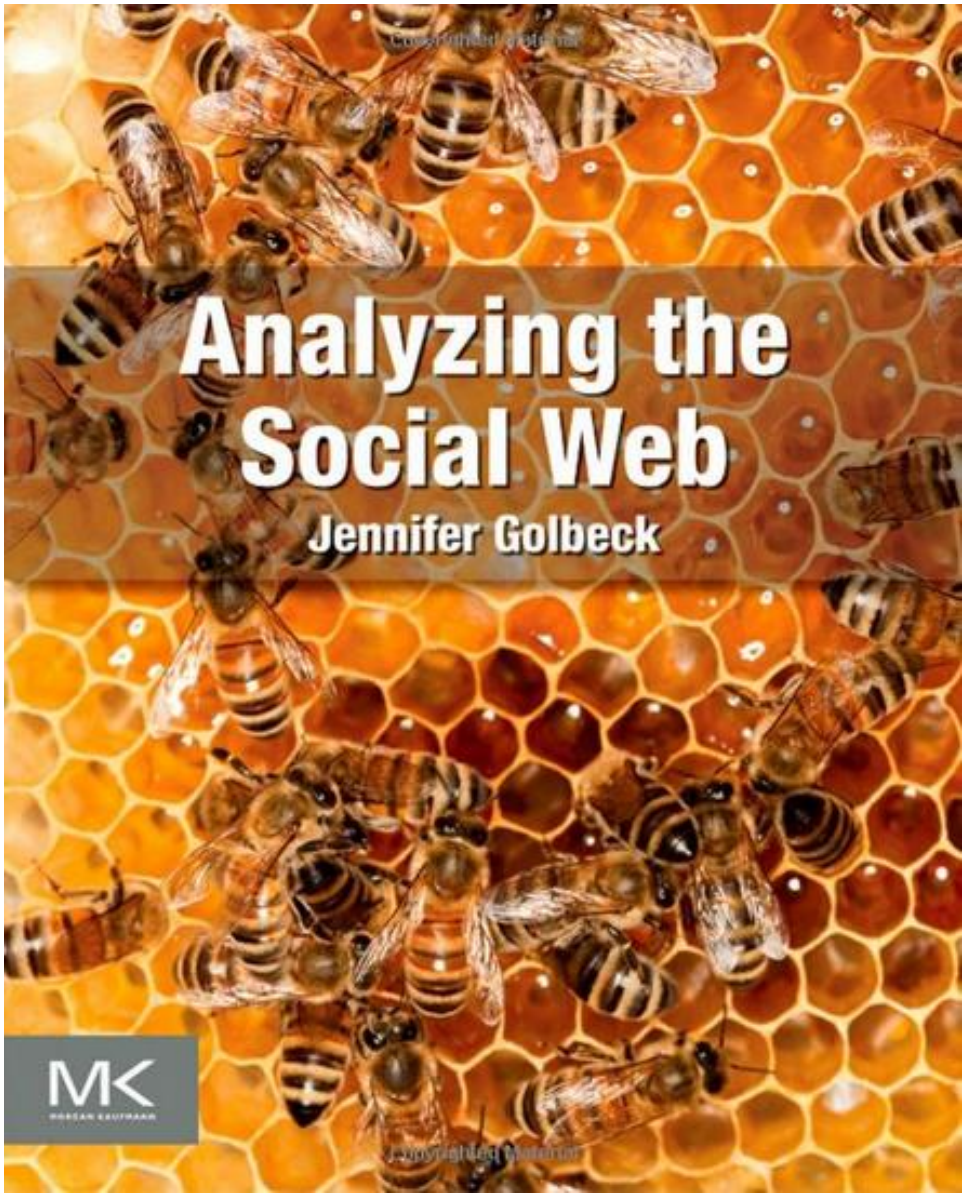
Yves Hilpisch, Python for Finance: Analyze Big Financial Data, O'Reilly, 2014



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

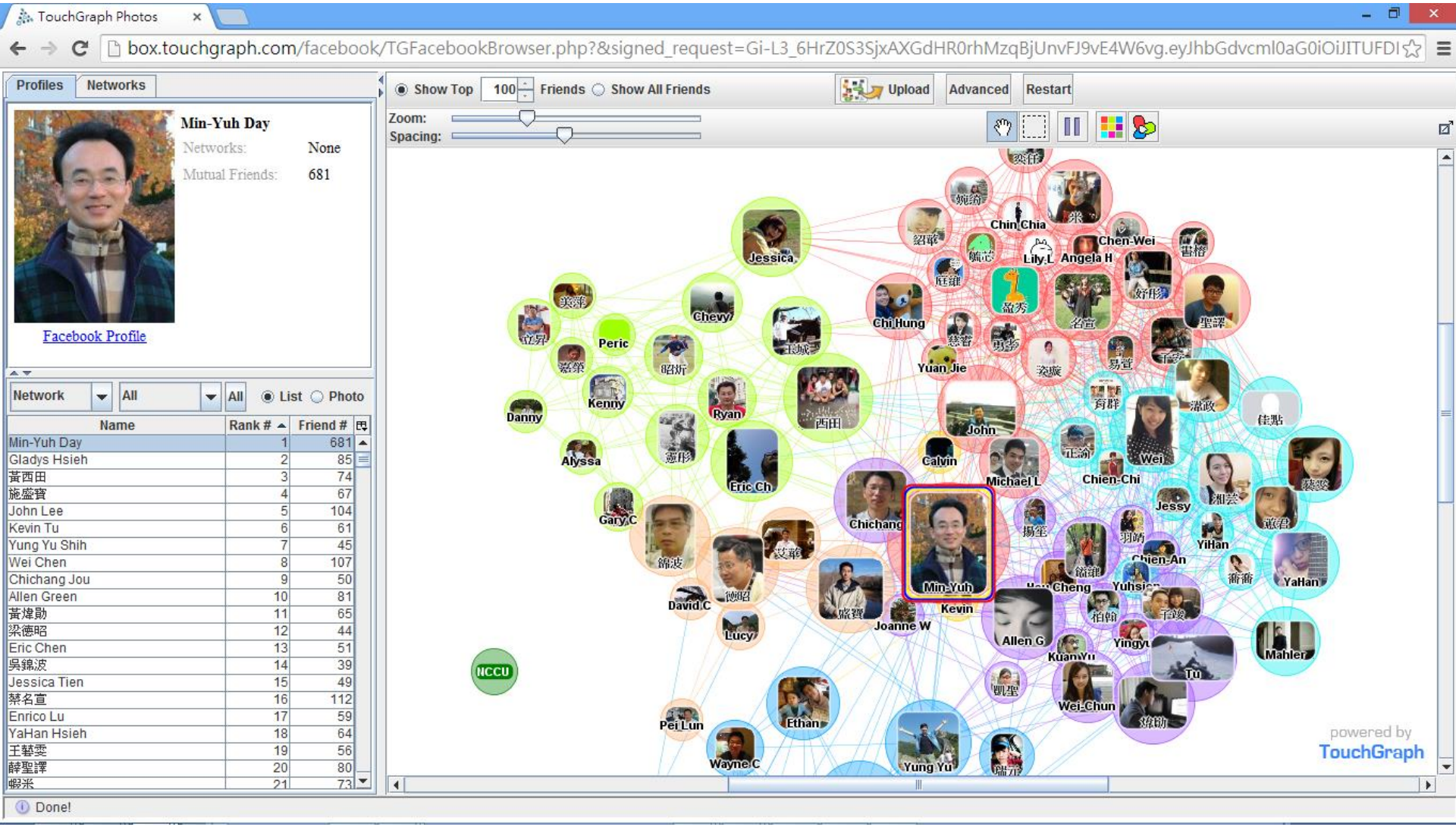
Analyzing the Social Web:
Social Network Analysis

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

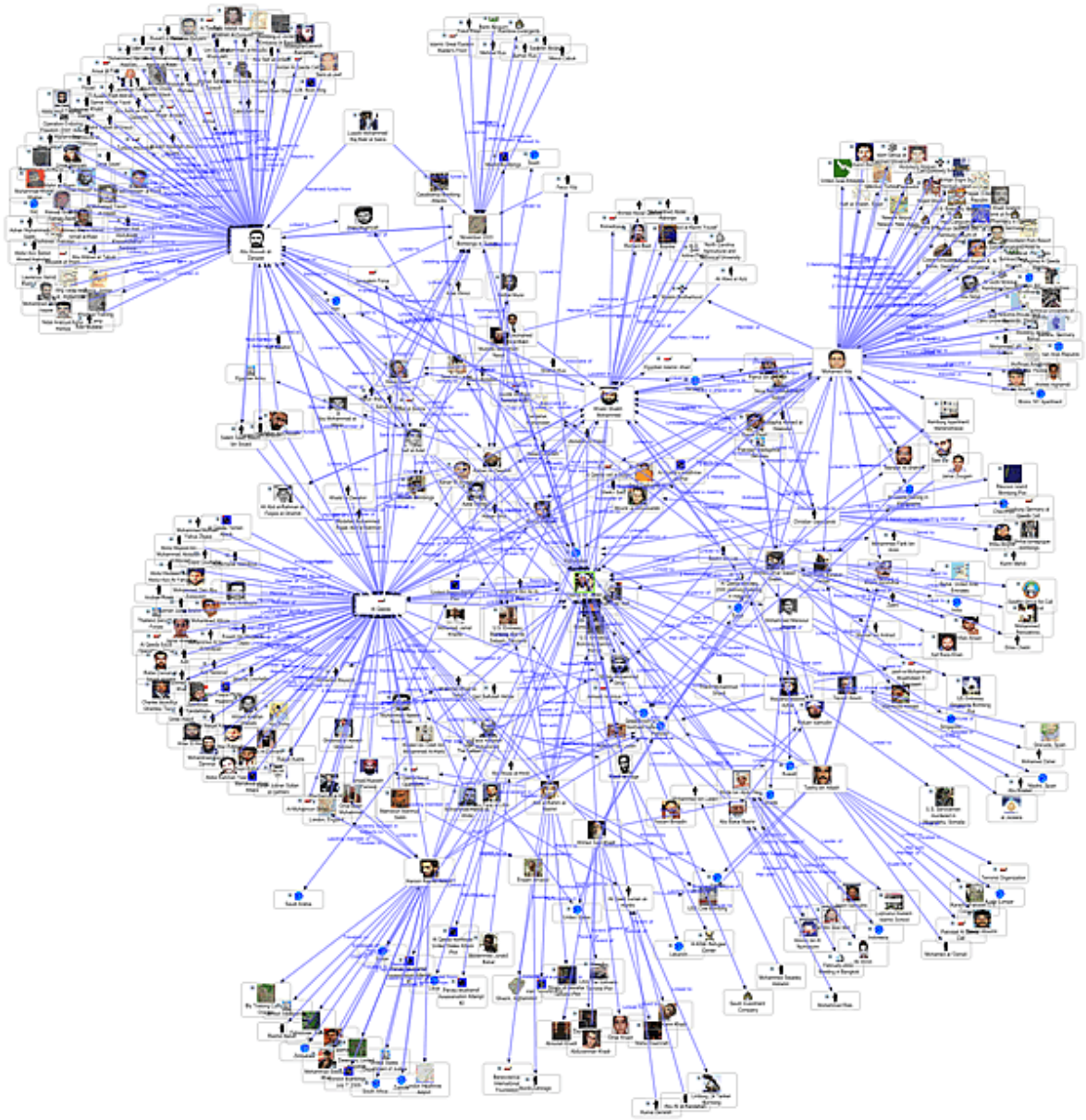


Source: http://www.amazon.com/Analyzing-Social-Web-Jennifer-Golbeck/dp/0124055311

Social Network Analysis (SNA)
Facebook TouchGraph



Social Network Analysis



Source: http://www.fmsasg.com/SocialNetworkAnalysis/

Social Network Analysis

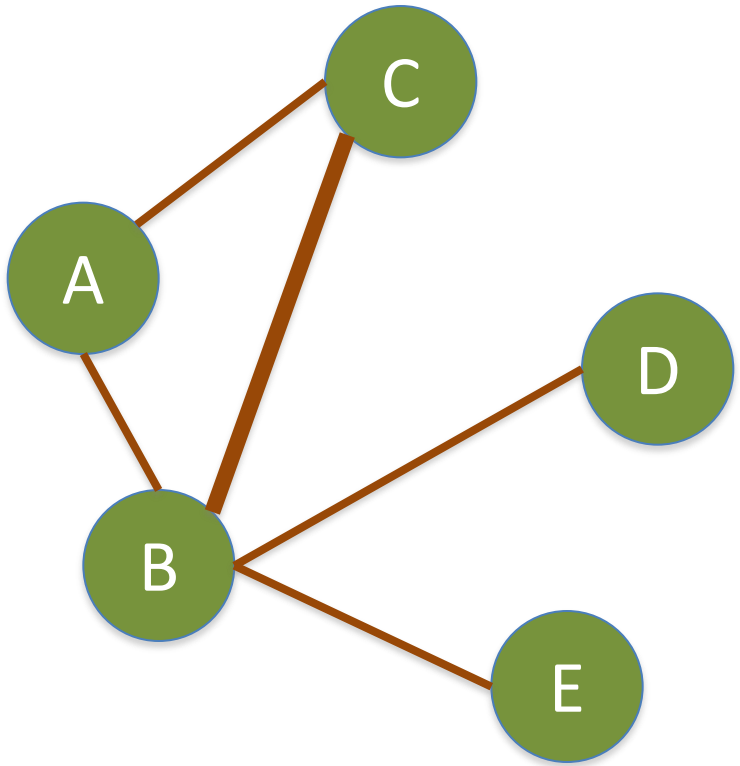
- A **social network** is a social structure of people, related (directly or indirectly) to each other through a common relation or interest
- **Social network analysis (SNA)** is the study of social networks to understand their structure and behavior

Social Network Analysis (SNA)

Centrality

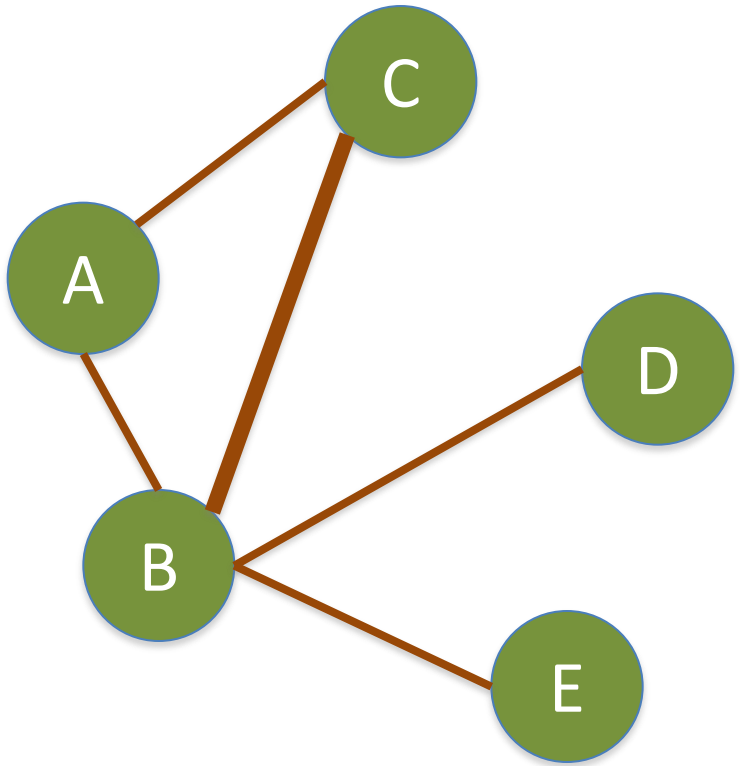
Prestige

Degree



Source: https://www.youtube.com/watch?v=89mxOdwPfxA

Degree



A: 2

B: 4

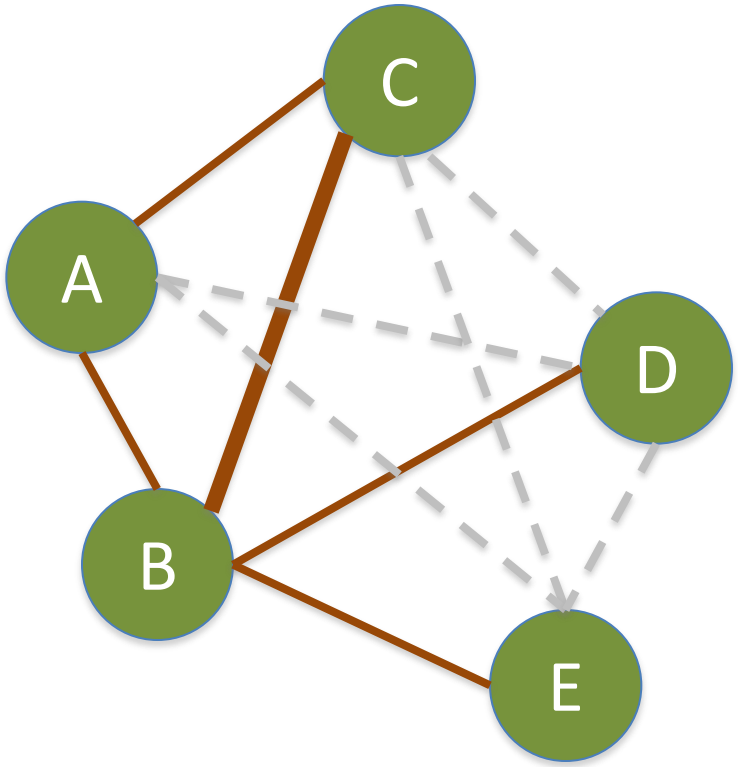
C: 2

D:1

E: 1

Source: https://www.youtube.com/watch?v=89mxOdwPfxA

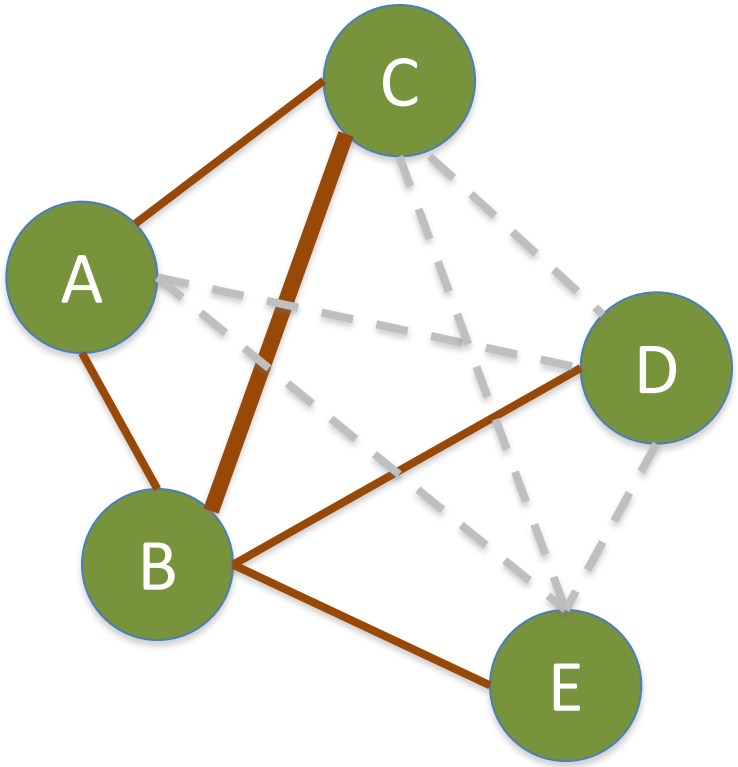
Density



Source: https://www.youtube.com/watch?v=89mxOdwPfxA

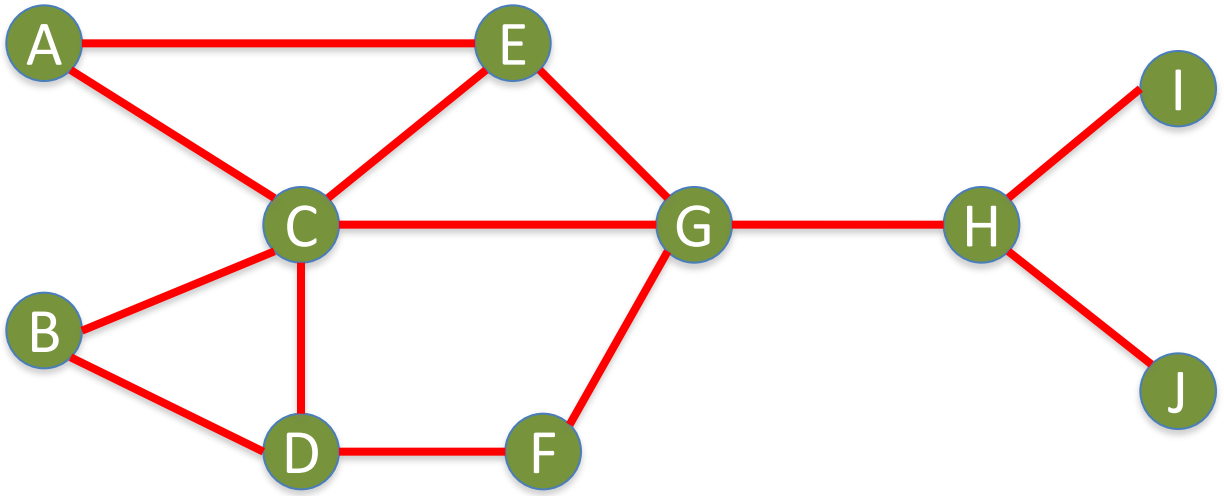
Density

Edges (Links): 5
Total Possible Edges: 10
Density: 5/10 = 0.5



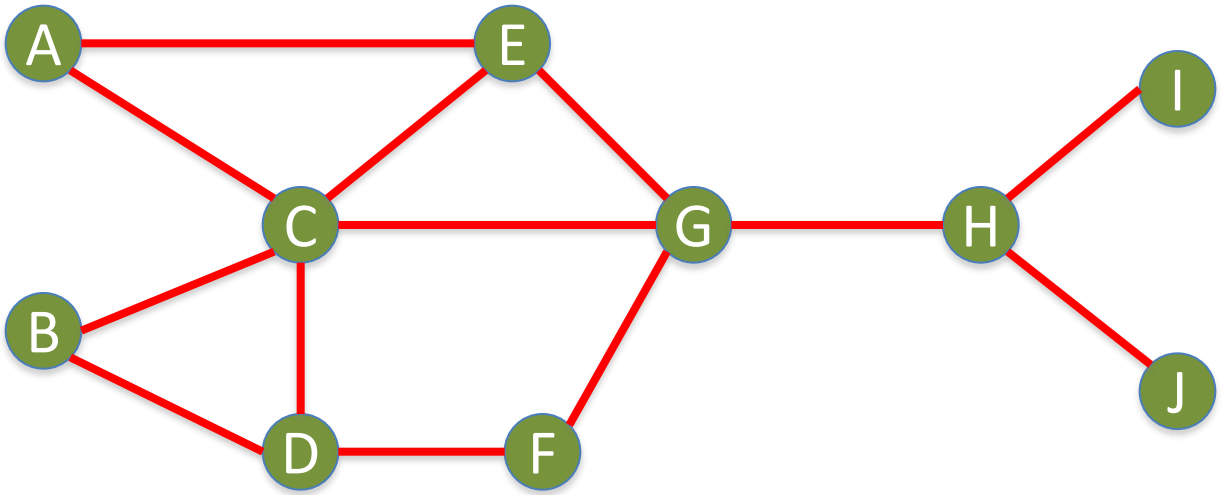
Source: https://www.youtube.com/watch?v=89mxOdwPfxA

Density



Nodes (n): 10
Edges (Links): 13
Total Possible Edges: (n * (n-1)) / 2 = (10 * 9) / 2 = 45
Density: 13/45 = 0.29

Which Node is Most Important?



Centrality

- Important or prominent actors are those that are linked or involved with other actors extensively.
- A person with extensive contacts (links) or communications with many other people in the organization is considered more important than a person with relatively fewer contacts.
- The links can also be called **ties**.
  A central actor is one involved in many ties.

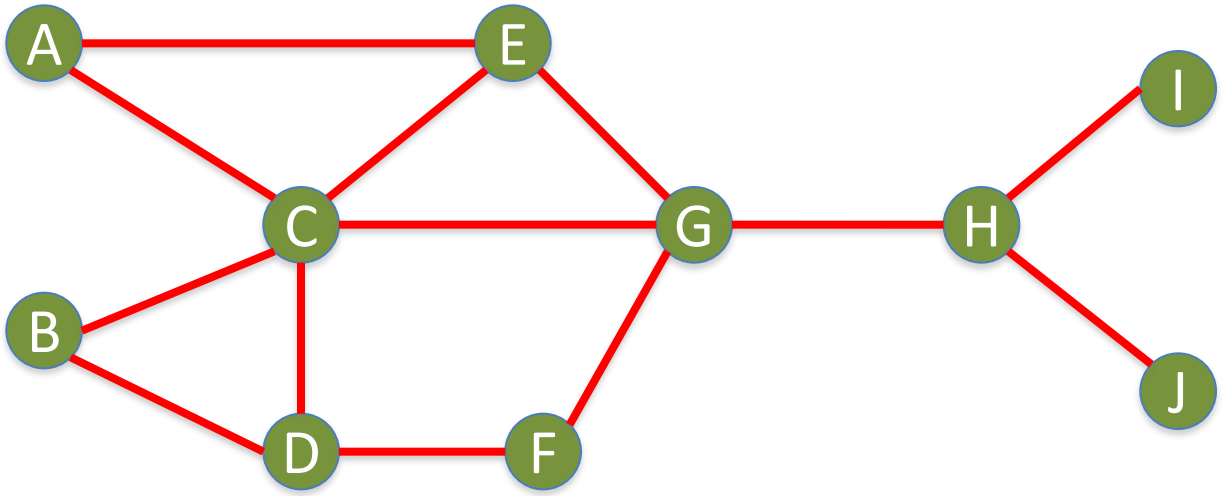
Source: Bing Liu (2011) , "Web Data Mining: Exploring Hyperlinks, Contents, and Usage Data"

Social Network Analysis (SNA)

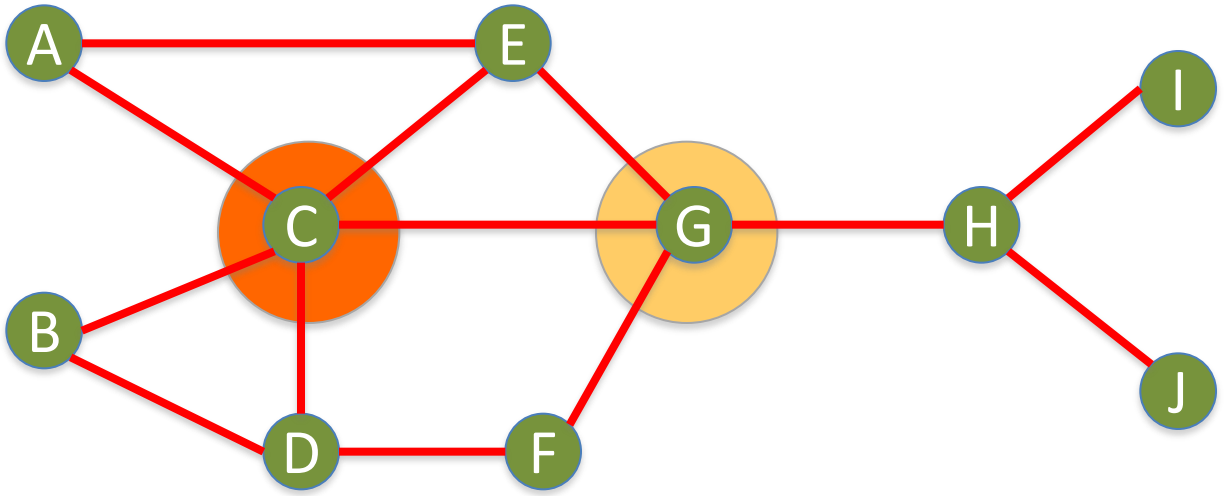
- Degree Centrality
- Betweenness Centrality
- Closeness Centrality

Degree Centrality

Social Network Analysis: Degree Centrality



Social Network Analysis: Degree Centrality



Node	Score	Standardized Score
A	2	2/10 = 0.2
B	2	2/10 = 0.2
C	**5**	**5/10 = 0.5**
D	3	3/10 = 0.3
E	3	3/10 = 0.3
F	2	2/10 = 0.2
G	**4**	**4/10 = 0.4**
H	3	3/10 = 0.3
I	1	1/10 = 0.1
J	1	1/10 = 0.1

Betweenness Centrality

Betweenness centrality:

Connectivity

Number of shortest paths going through the actor

Betweenness Centrality

$$C_B(i) = \sum_{j<k} g_{ik}(i) / g_{jk}$$

Where g_{jk} = the number of shortest paths connecting jk
$g_{jk}(i)$ = the number that actor i is on.

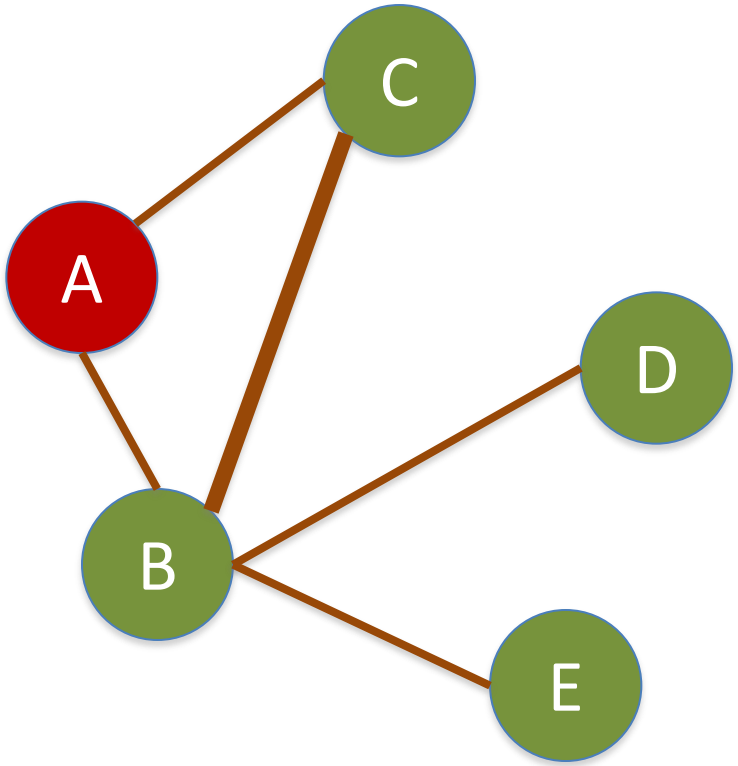
Normalized Betweenness Centrality

$$C'_B(i) = C_B(i) / [(n-1)(n-2)/2]$$

Number of pairs of vertices excluding the vertex itself

Source: https://www.youtube.com/watch?v=RXohUeNCJiU

Betweenness Centrality

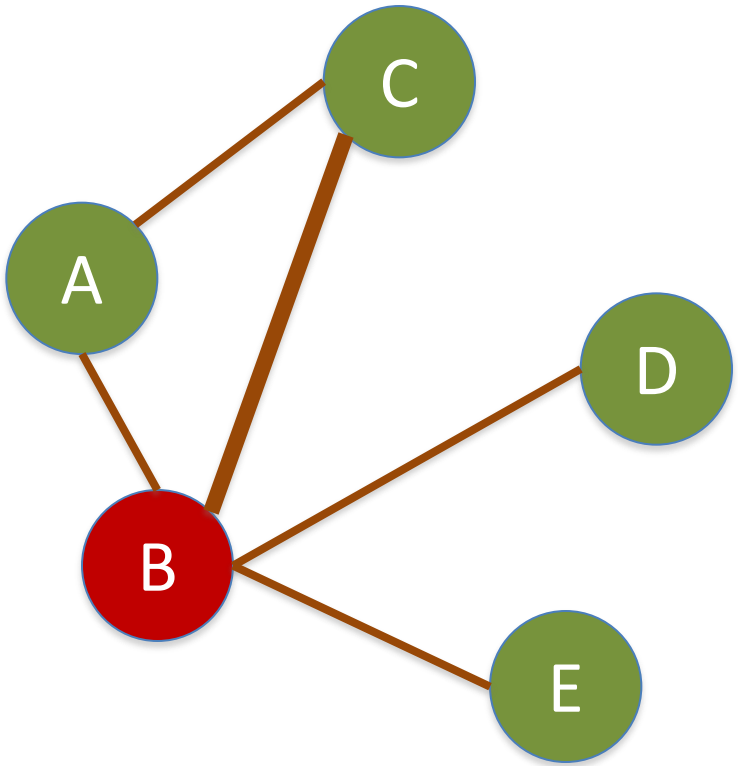


A:
B→C: 0/1 = 0
B→D: 0/1 = 0
B→E: 0/1 = 0
C→D: 0/1 = 0
C→E: 0/1 = 0
D→E: 0/1 = 0

Total: 0

A: Betweenness Centrality = 0

Betweenness Centrality

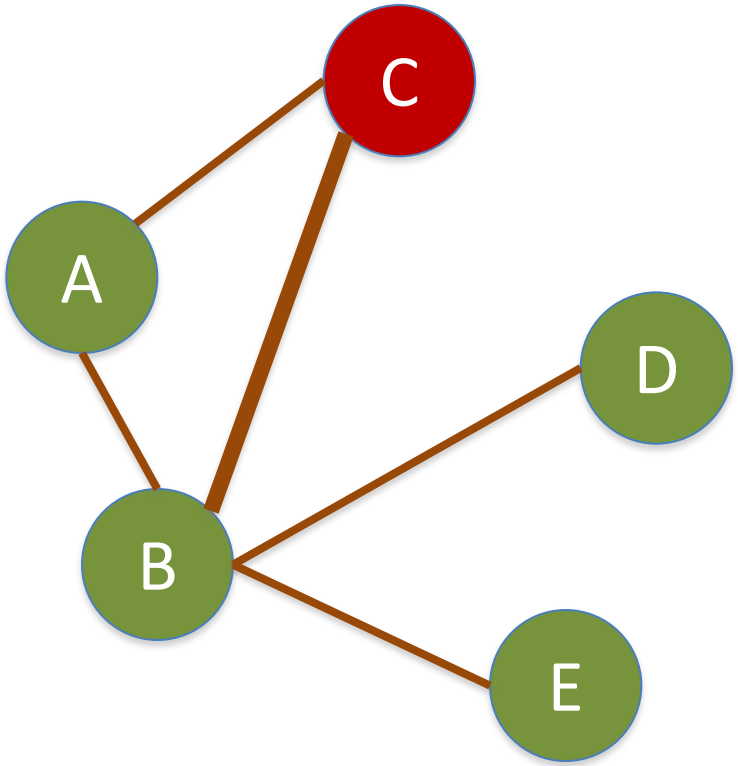


B:
A→C: 0/1 = 0
A→D: 1/1 = 1
A→E: 1/1 = 1
C→D: 1/1 = 1
C→E: 1/1 = 1
D→E: 1/1 = 1

Total: 5

B: Betweenness Centrality = 5

Betweenness Centrality

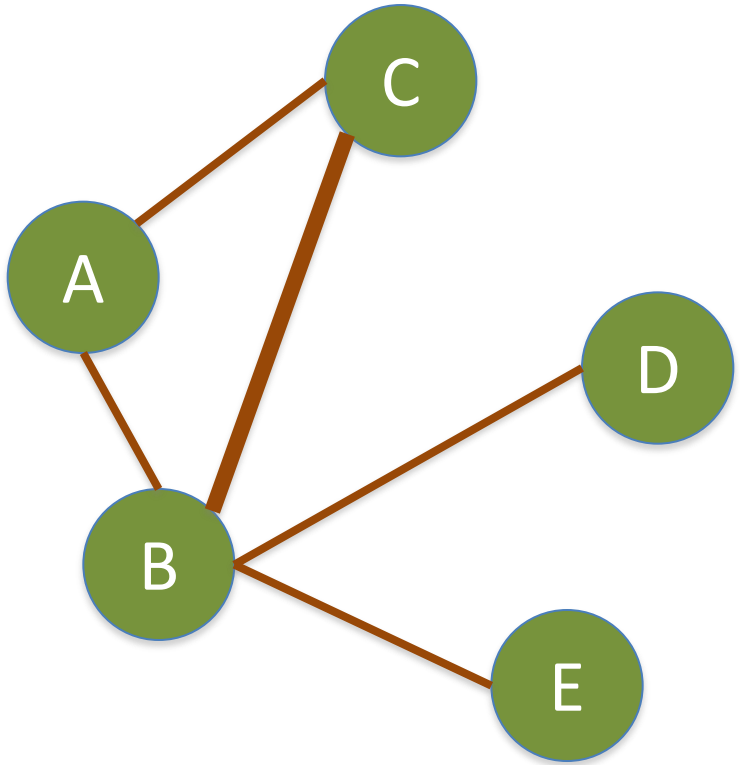


C:
A→B: 0/1 = 0
A→D: 0/1 = 0
A→E: 0/1 = 0
B→D: 0/1 = 0
B→E: 0/1 = 0
D→E: 0/1 = 0

Total: 0

C: Betweenness Centrality = 0

Betweenness Centrality



A: 0

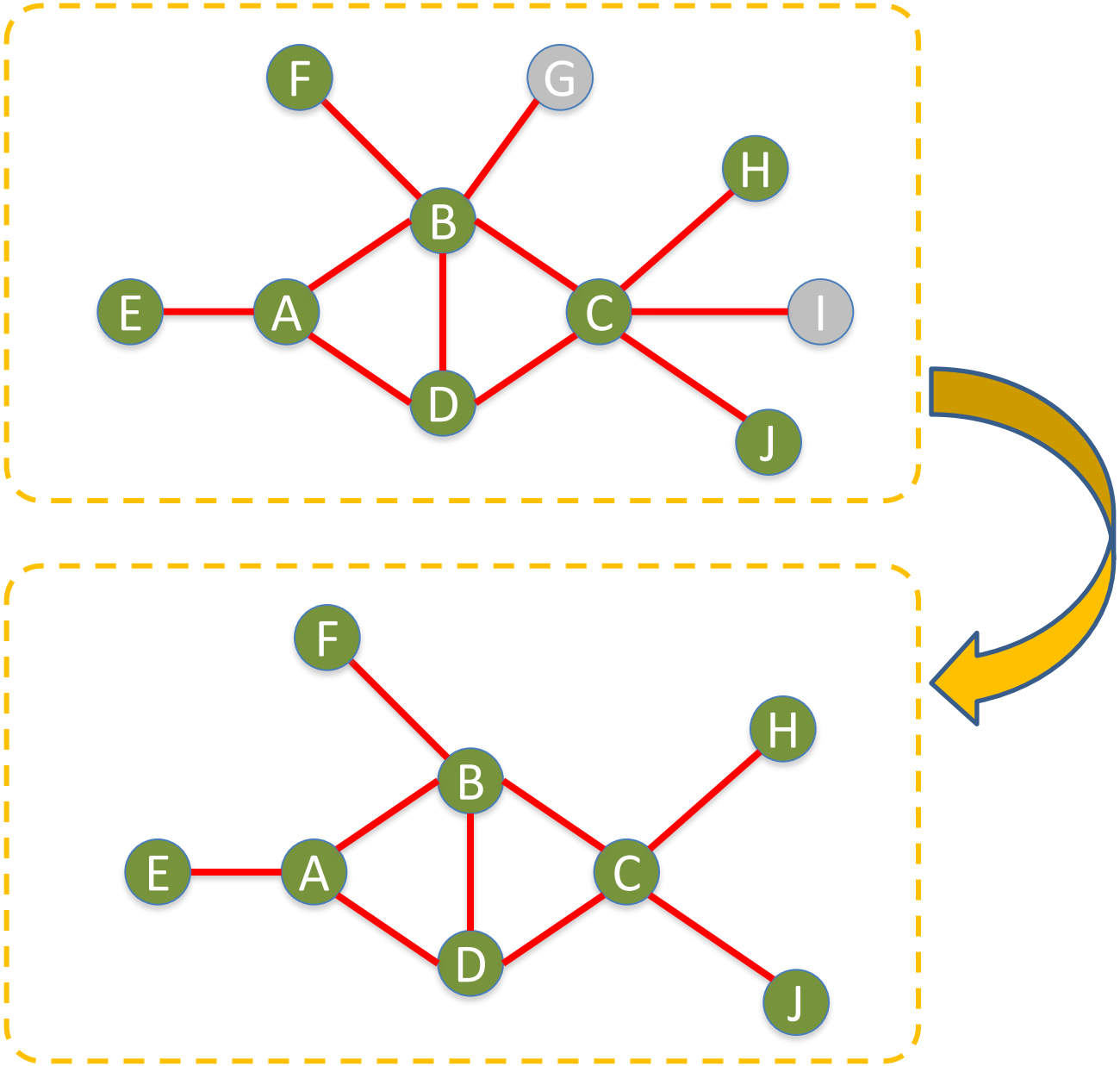
B: 5

C: 0

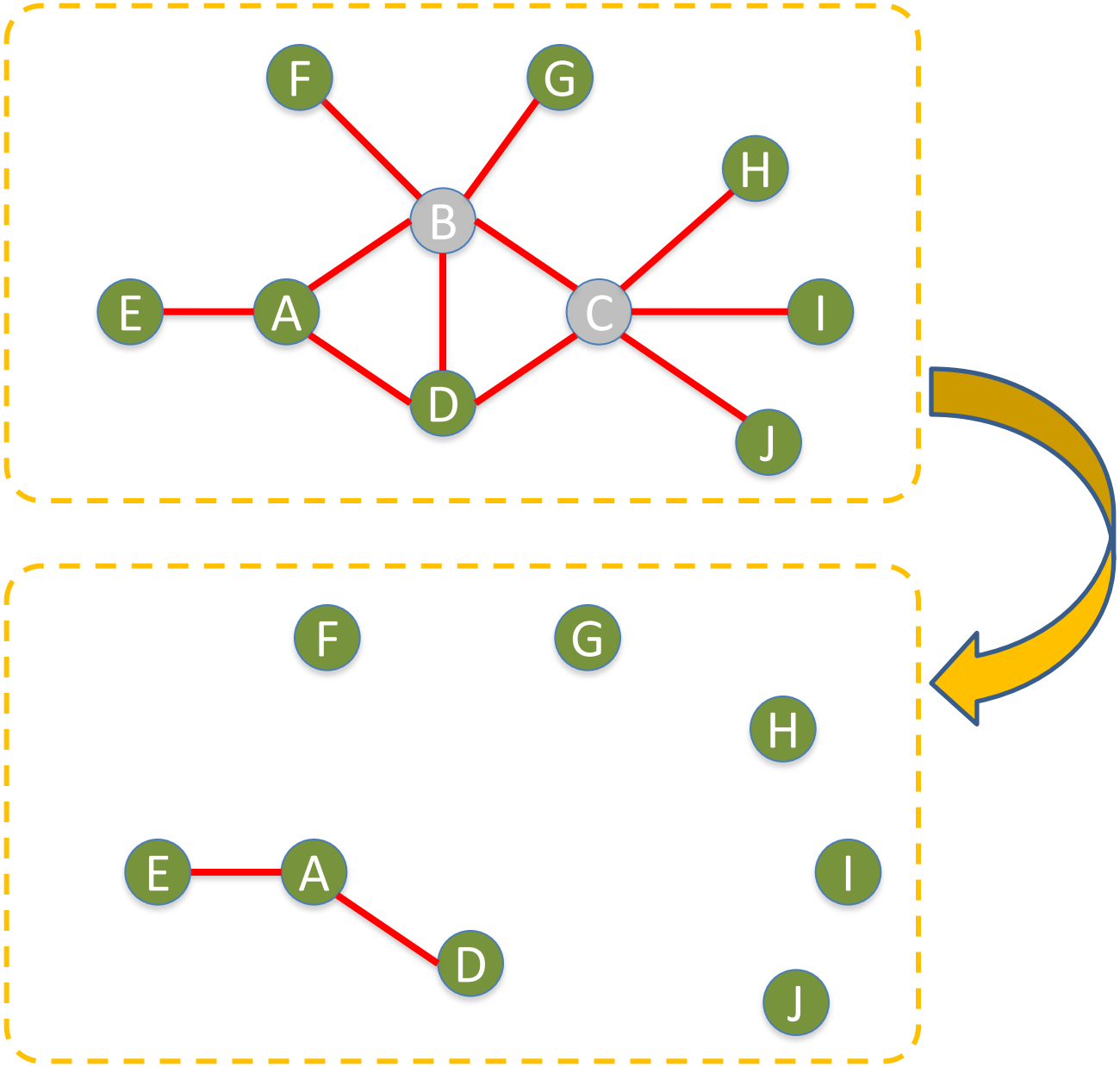
D: 0

E: 0

Which Node is Most Important?

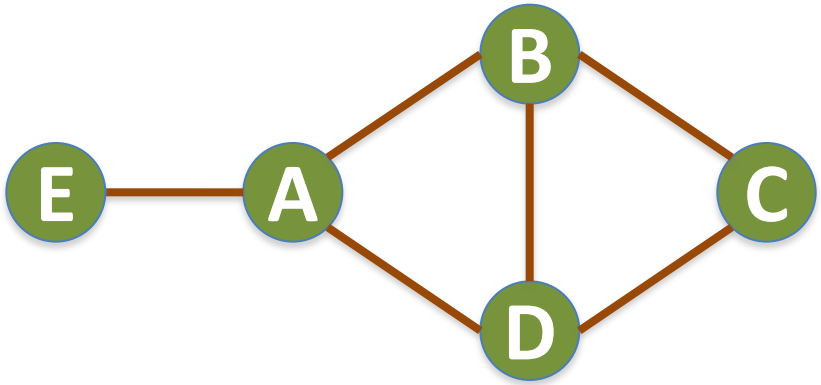


Which Node is Most Important?

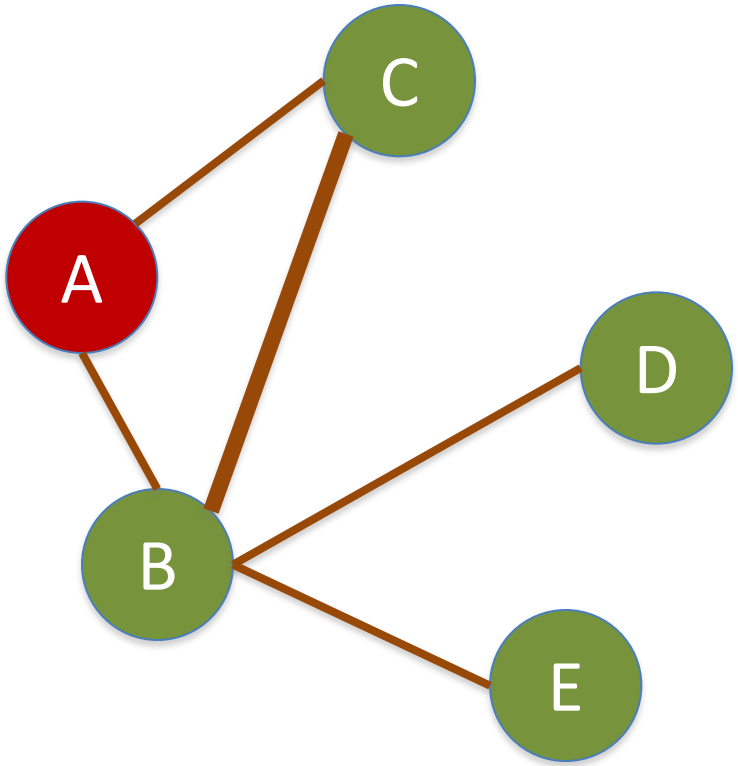


Betweenness Centrality

$$C_B(i) = \sum_{j<k} g_{ik}(i) / g_{jk}$$



Betweenness Centrality



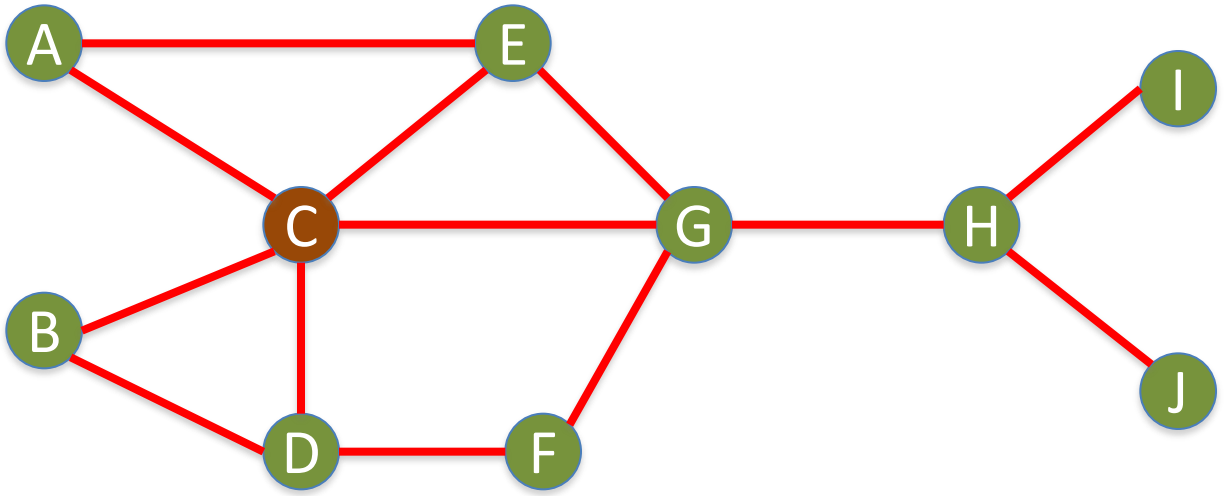
A:
B→C: 0/1 = 0
B→D: 0/1 = 0
B→E: 0/1 = 0
C→D: 0/1 = 0
C→E: 0/1 = 0
D→E: 0/1 = 0

Total: 0

A: Betweenness Centrality = 0

Closeness Centrality

Social Network Analysis: Closeness Centrality

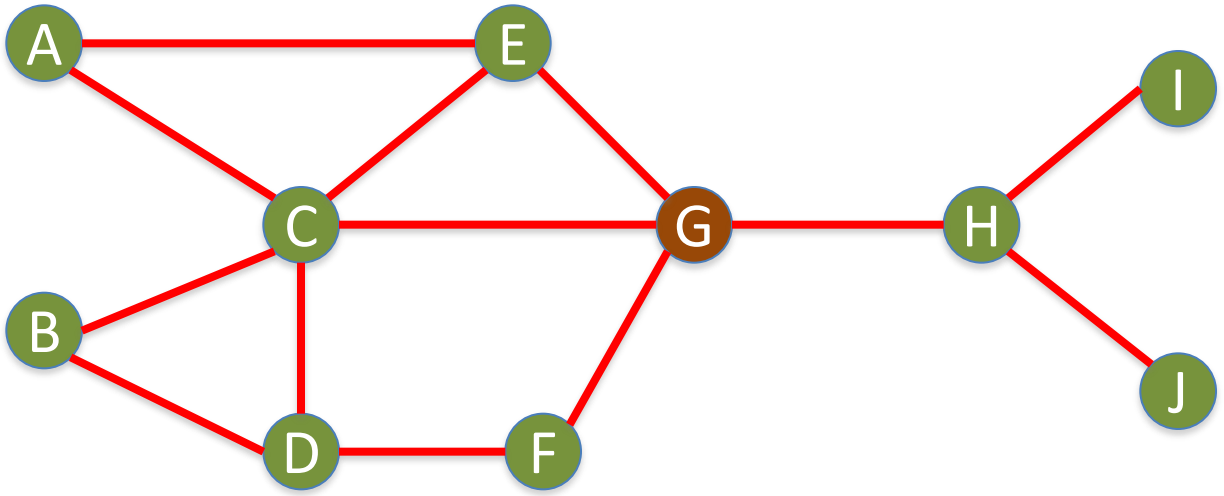


C→A: 1
C→B: 1
C→D: 1
C→E: 1
C→F: 2
C→G: 1
C→H: 2
C→I: 3
C→J: 3

Total=15

C: Closeness Centrality = 15/9 = 1.67

Social Network Analysis: Closeness Centrality

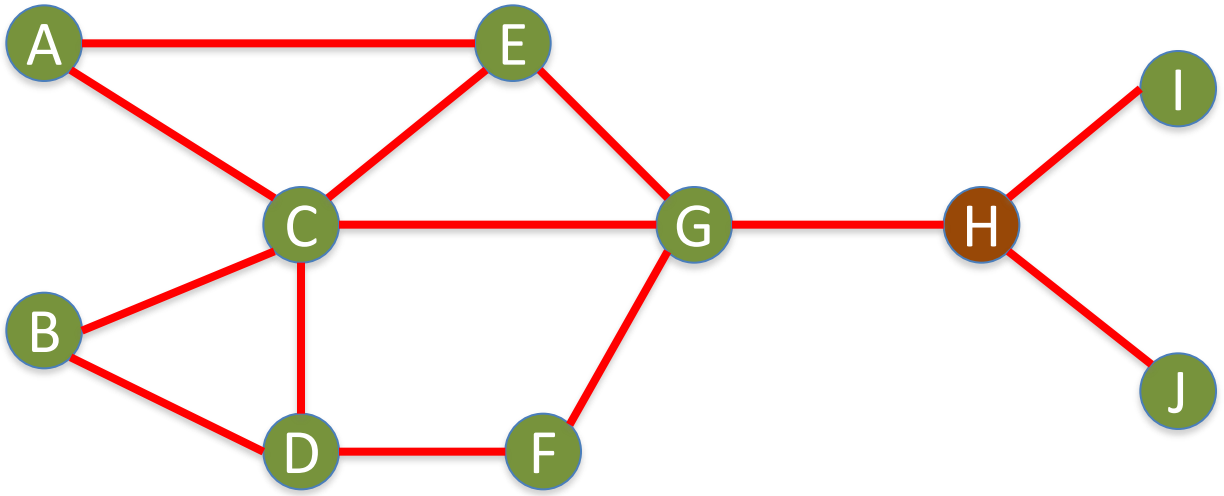


G→A: 2
G→B: 2
G→C: 1
G→D: 2
G→E: 1
G→F: 1
G→H: 1
G→I: 2
G→J: 2

Total=14

G: Closeness Centrality = 14/9 = 1.56

Social Network Analysis: Closeness Centrality

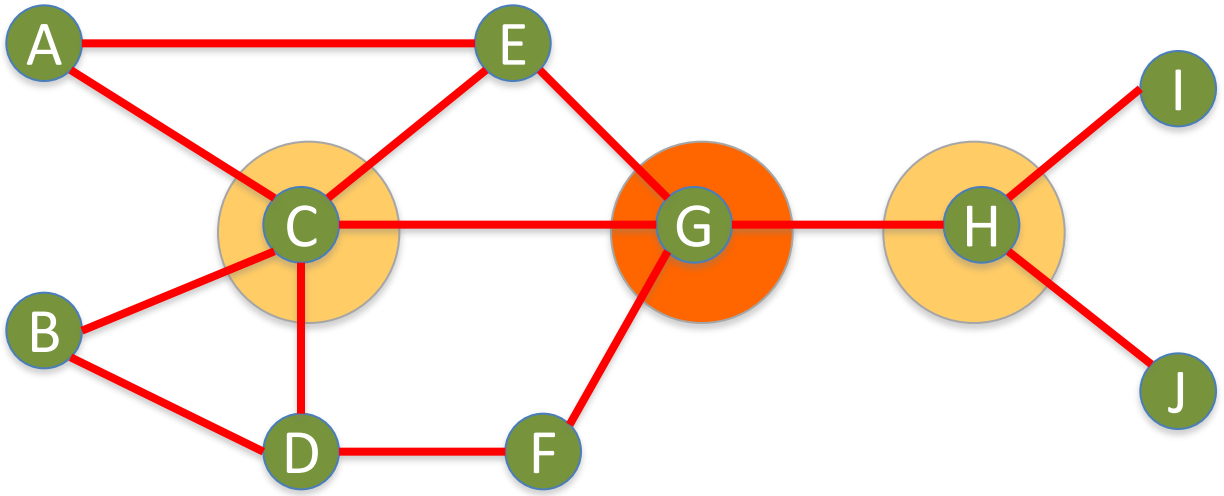


H→A: 3
H→B: 3
H→C: 2
H→D: 2
H→E: 2
H→F: 2
H→G: 1
H→I: 1
H→J: 1

Total=17

H: Closeness Centrality = 17/9 = 1.89

Social Network Analysis: Closeness Centrality

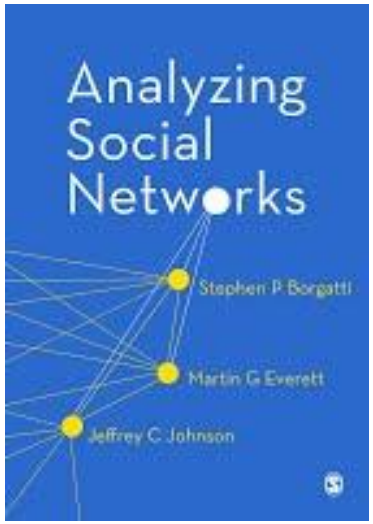


G: Closeness Centrality = 14/9 = 1.56 1

C: Closeness Centrality = 15/9 = 1.67 2

H: Closeness Centrality = 17/9 = 1.89 3

Social Network Analysis (SNA) Tools



- **UCINet**
- **Pajek**

Application of SNA

Social Network Analysis of Research Collaboration in Information Reuse and Integration

Source: Min-Yuh Day, Sheng-Pao Shih, Weide Chang (2011),
"Social Network Analysis of Research Collaboration in Information Reuse and Integration"

Example of SNA Data Source

dblp
computer science bibliography

home | browse | search | about

IRI 2010: Las Vegas, NV, USA

- **Proceedings of the IEEE International Conference on Information Reuse and Integration, IRI 2010, 4-6 August 2010, Las Vegas, Nevada, USA.** IEEE Systems, Man, and Cybernetics Society 2010
- Reda Alhajj, James B. D. Joshi, Mei-Ling Shyu: **Message from Program Co-Chairs.** 1
- Stuart Harvey Rubin, Shu-Ching Chen: **Forward.** 1
- Lotfi A. Zadeh: **Precisiation of meaning - toward computation with natural language.** 1-4
- Reda Alhajj, Shu-Ching Chen, Gongzhu Hu, James B. D. Joshi, Gordon K. Lee, Stuart Harvey Rubin, Mei-Ling Shyu, Lotfi A. Zadeh: **Panel title: Critical need for funding of basic and applied research in large-scale computing.** 1

Automation, Integration and Reuse across Various Apps

- László István Etesi, André Csillaghy, Lin-Ching Chang: **A message-based interoperability framework with application to astrophysics.** 1-6
- Awny Alnusair, Tian Zhao, Eric Bodden: **Effective API navigation and reuse.** 7-12
- Manabu Ohta, Ryohei Inoue, Atsuhiro Takasu: **Empirical evaluation of active sampling for CRF-based analysis of pages.** 13-18
- Qunzhi Zhou, Viktor K. Prasanna: **Workflow management of simulation based computation processes in transportation domain.** 19-24

Source: http://www.informatik.uni-trier.de/~ley/db/conf/iri/iri2010.html

Research Question

- RQ1: What are the scientific collaboration patterns in the IRI research community?

- RQ2: Who are the prominent researchers in the IRI community?

Source: Min-Yuh Day, Sheng-Pao Shih, Weide Chang (2011),
"Social Network Analysis of Research Collaboration in Information Reuse and Integration"

Methodology

- Developed a simple web focused crawler program to download literature information about all IRI papers published between 2003 and 2010 from IEEE Xplore and DBLP.
  - 767 paper
  - 1599 distinct author
- Developed a program to convert the list of coauthors into the format of a network file which can be readable by social network analysis software.
- *UCINet* and *Pajek* were used in this study for the social network analysis.

Source: Min-Yuh Day, Sheng-Pao Shih, Weide Chang (2011),
"Social Network Analysis of Research Collaboration in Information Reuse and Integration"

Top10 prolific authors (IRI 2003-2010)

1. Stuart Harvey Rubin
2. Taghi M. Khoshgoftaar
3. Shu-Ching Chen
4. Mei-Ling Shyu
5. Mohamed E. Fayad
6. Reda Alhajj
7. Du Zhang
8. Wen-Lian Hsu
9. Jason Van Hulse
10. Min-Yuh Day

Source: Min-Yuh Day, Sheng-Pao Shih, Weide Chang (2011),
"Social Network Analysis of Research Collaboration in Information Reuse and Integration"

Data Analysis and Discussion

- **Closeness Centrality**
  - Collaborated widely
- **Betweenness Centrality**
  - Collaborated diversely
- **Degree Centrality**
  - Collaborated frequently
- **Visualization of Social Network Analysis**
  - Insight into the structural characteristics of research collaboration networks

Source: Min-Yuh Day, Sheng-Pao Shih, Weide Chang (2011),
"Social Network Analysis of Research Collaboration in Information Reuse and Integration"

Top 20 authors with the highest closeness scores

Rank	ID	Closeness	Author
1	3	0.024675	Shu-Ching Chen
2	1	0.022830	Stuart Harvey Rubin
3	4	0.022207	Mei-Ling Shyu
4	6	0.020013	Reda Alhajj
5	61	0.019700	Na Zhao
6	260	0.018936	Min Chen
7	151	0.018230	Gordon K. Lee
8	19	0.017962	Chengcui Zhang
9	1043	0.017962	Isai Michel Lombera
10	1027	0.017962	Michael Armella
11	443	0.017448	James B. Law
12	157	0.017082	Keqi Zhang
13	253	0.016731	Shahid Hamid
14	1038	0.016618	Walter Z. Tang
15	959	0.016285	Chengjun Zhan
16	957	0.016285	Lin Luo
17	956	0.016285	Guo Chen
18	955	0.016285	Xin Huang
19	943	0.016285	Sneh Gulati
20	960	0.016071	Sheng-Tun Li

Source: Min-Yuh Day, Sheng-Pao Shih, Weide Chang (2011),
"Social Network Analysis of Research Collaboration in Information Reuse and Integration"

Top 20 authors with the highest betweeness scores

Rank	ID	Betweenness	Author
1	1	0.000752	Stuart Harvey Rubin
2	3	0.000741	Shu-Ching Chen
3	2	0.000406	Taghi M. Khoshgoftaar
4	66	0.000385	Xingquan Zhu
5	4	0.000376	Mei-Ling Shyu
6	6	0.000296	Reda Alhajj
7	65	0.000256	Xindong Wu
8	19	0.000194	Chengcui Zhang
9	39	0.000185	Wei Dai
10	15	0.000107	Narayan C. Debnath
11	31	0.000094	Qianhui Althea Liang
12	151	0.000094	Gordon K. Lee
13	7	0.000085	Du Zhang
14	30	0.000072	Baowen Xu
15	41	0.000067	Hongji Yang
16	270	0.000060	Zhiwei Xu
17	5	0.000043	Mohamed E. Fayad
18	110	0.000042	Abhijit S. Pandya
19	106	0.000042	Sam Hsu
20	8	0.000042	Wen-Lian Hsu

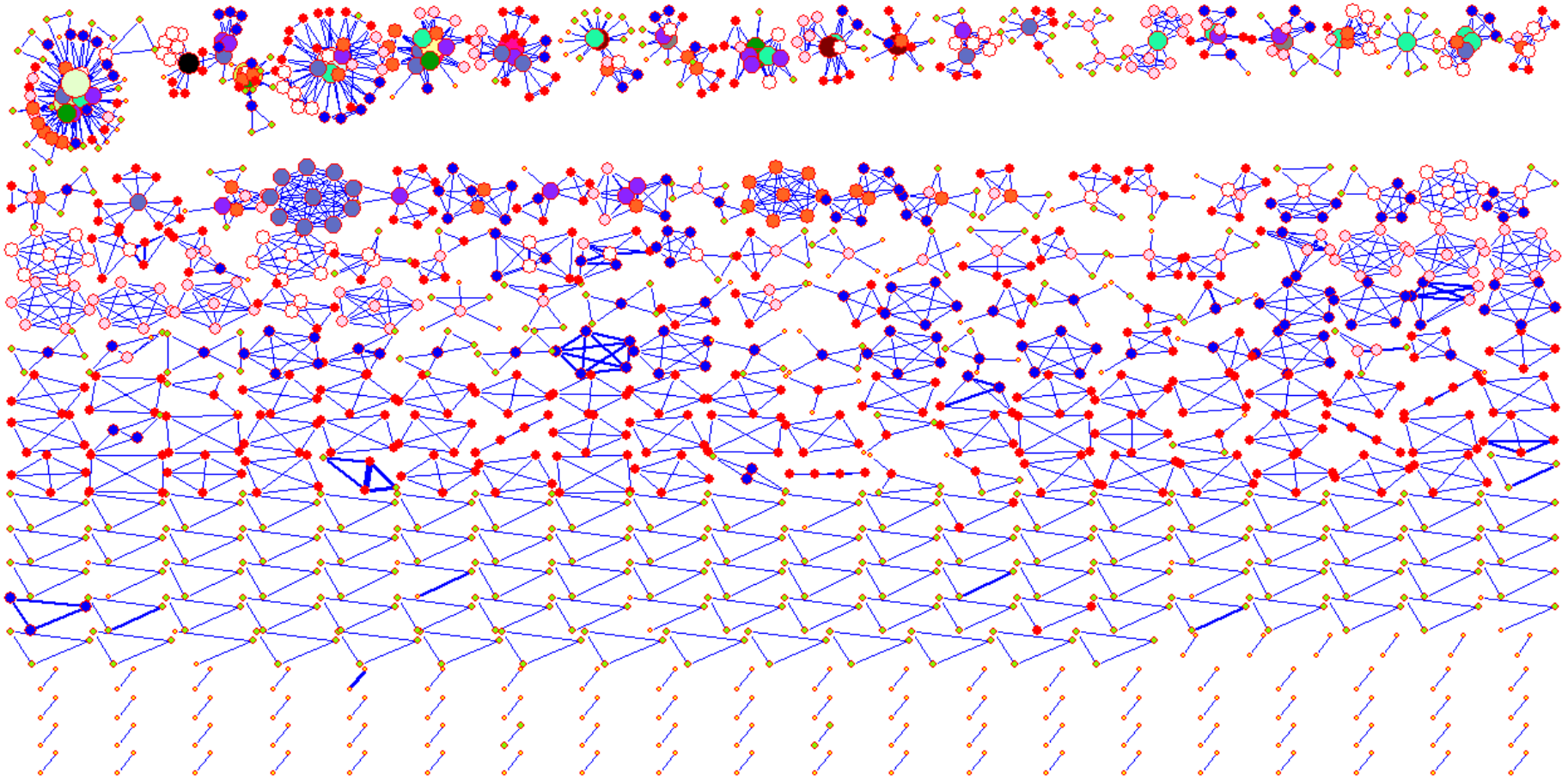
Source: Min-Yuh Day, Sheng-Pao Shih, Weide Chang (2011),
"Social Network Analysis of Research Collaboration in Information Reuse and Integration"

Top 20 authors with the highest degree scores

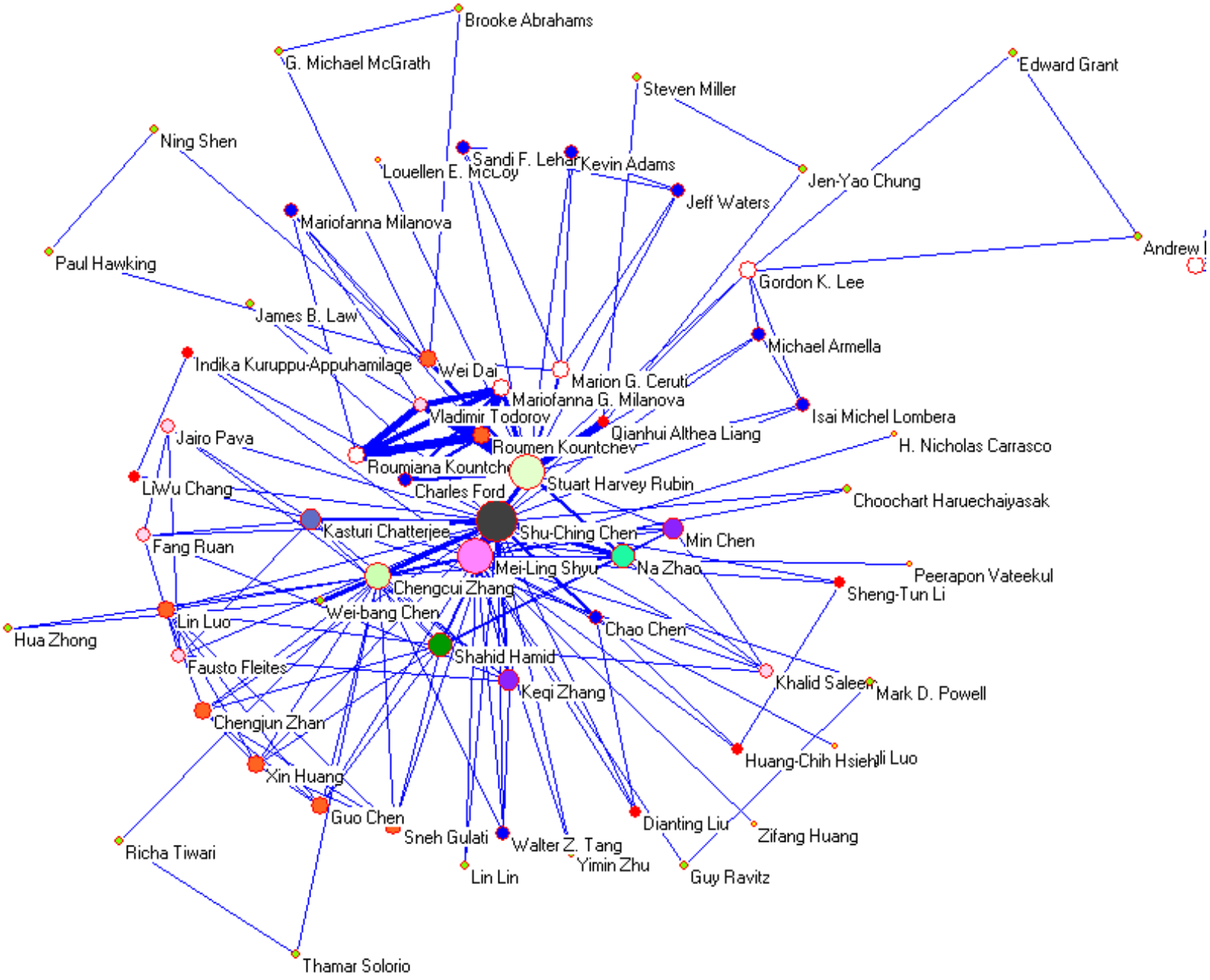
Rank	ID	Degree	Author
1	3	0.035044	Shu-Ching Chen
2	1	0.034418	Stuart Harvey Rubin
3	2	0.030663	Taghi M. Khoshgoftaar
4	6	0.028786	Reda Alhajj
5	8	0.028786	Wen-Lian Hsu
6	10	0.024406	Min-Yuh Day
7	4	0.022528	Mei-Ling Shyu
8	17	0.021277	Richard Tzong-Han Tsai
9	14	0.017522	Eduardo Santana de Almeida
10	16	0.017522	Roumen Kountchev
11	40	0.016896	Hong-Jie Dai
12	15	0.015645	Narayan C. Debnath
13	9	0.015019	Jason Van Hulse
14	25	0.013767	Roumiana Kountcheva
15	28	0.013141	Silvio Romero de Lemos Meira
16	24	0.013141	Vladimir Todorov
17	23	0.013141	Mariofanna G. Milanova
18	5	0.013141	Mohamed E. Fayad
19	19	0.012516	Chengcui Zhang
20	18	0.011890	Waleed W. Smari

Source: Min-Yuh Day, Sheng-Pao Shih, Weide Chang (2011),
"Social Network Analysis of Research Collaboration in Information Reuse and Integration"

Visualization of IRI (IEEE IRI 2003-2010) co-authorship network (global view)

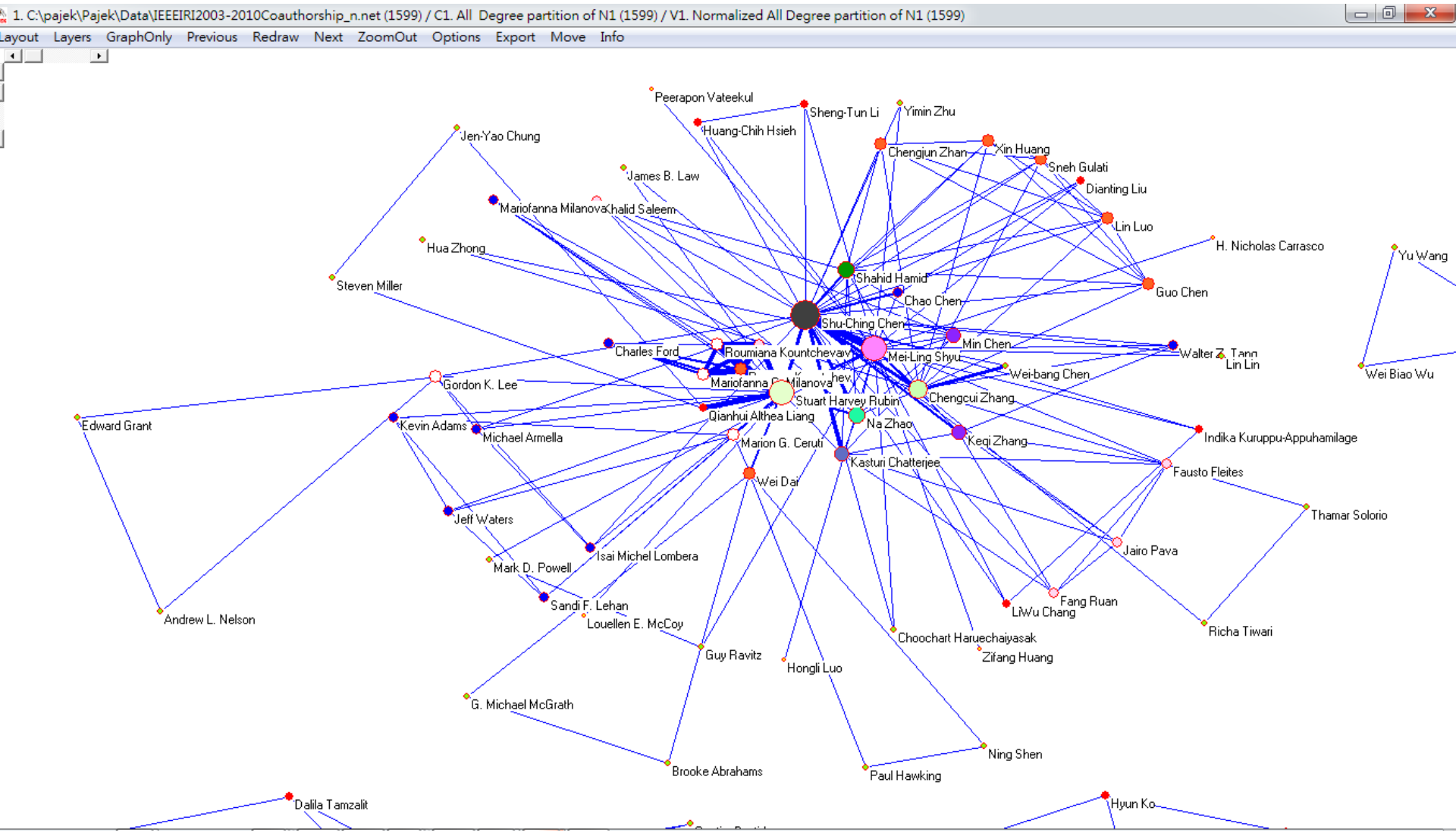


Source: Min-Yuh Day, Sheng-Pao Shih, Weide Chang (2011),
"Social Network Analysis of Research Collaboration in Information Reuse and Integration"



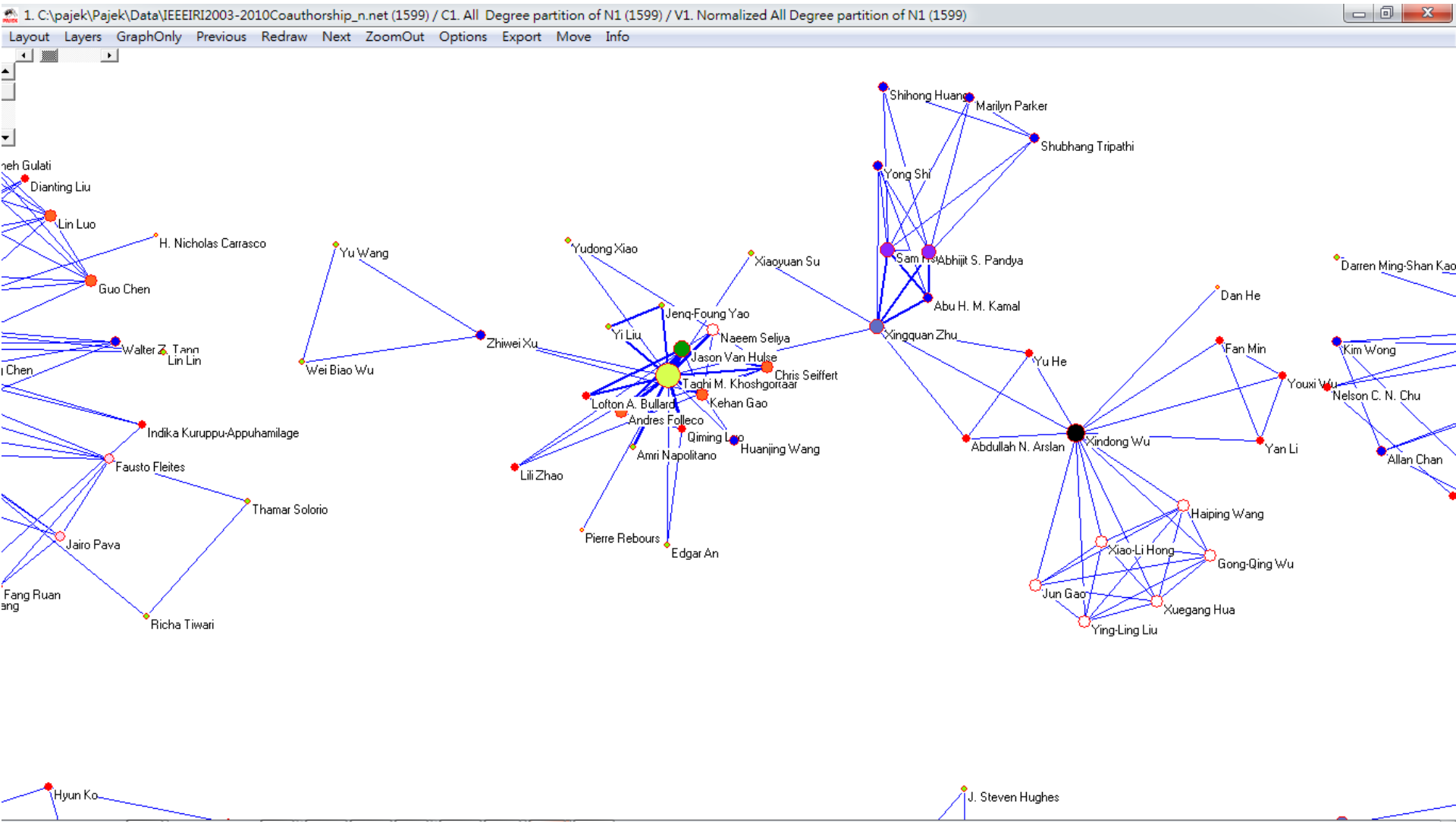
Source: Min-Yuh Day, Sheng-Pao Shih, Weide Chang (2011),
"Social Network Analysis of Research Collaboration in Information Reuse and Integration"

Visualization of Social Network Analysis



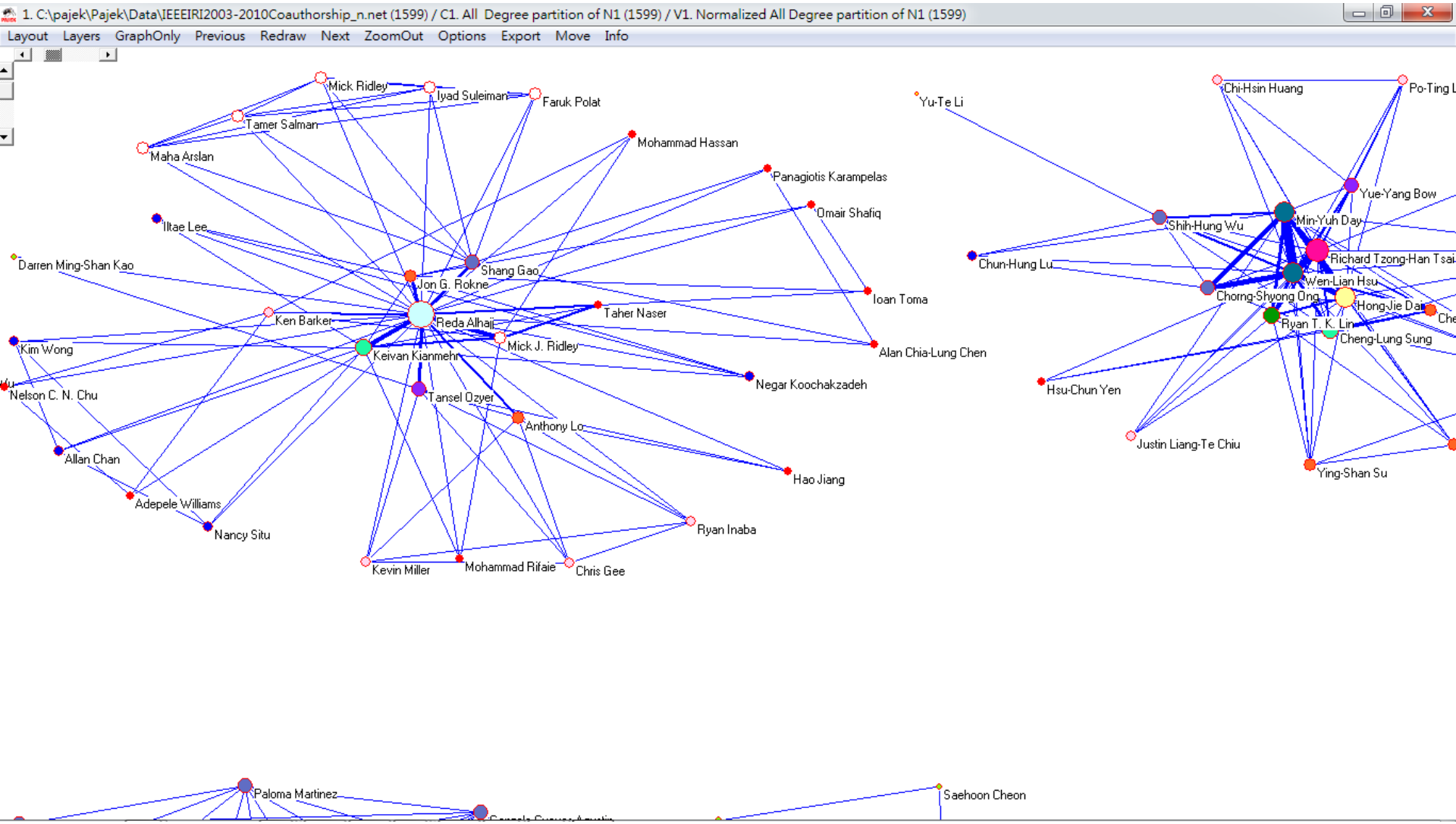
Source: Min-Yuh Day, Sheng-Pao Shih, Weide Chang (2011),
"Social Network Analysis of Research Collaboration in Information Reuse and Integration"

Visualization of Social Network Analysis



Source: Min-Yuh Day, Sheng-Pao Shih, Weide Chang (2011),
"Social Network Analysis of Research Collaboration in Information Reuse and Integration"

Visualization of Social Network Analysis



Source: Min-Yuh Day, Sheng-Pao Shih, Weide Chang (2011),
"Social Network Analysis of Research Collaboration in Information Reuse and Integration"

NTCIR 12
QALab-2 Task

http://research.nii.ac.jp/qalab/

Overview of NTCIR Evaluation Activities

NTCIR

NII Testbeds and Community for Information access Research

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

NII:
National Institute of Informatics

http://www.nii.ac.jp/en/

NTCIR

NII Testbeds and Community for Information access Research

NTCIR

Research Infrastructure for Evaluating Information Access

- A series of evaluation workshops designed to enhance research in information-access technologies by providing an infrastructure for large-scale evaluations.
- Data sets, evaluation methodologies, forum

Source: Kando et al., 2013

NTCIR

- **Project started in late 1997**
  - 18 months Cycle

Source: Kando et al., 2013

NTCIR

NII Testbeds and Community for Information access Research

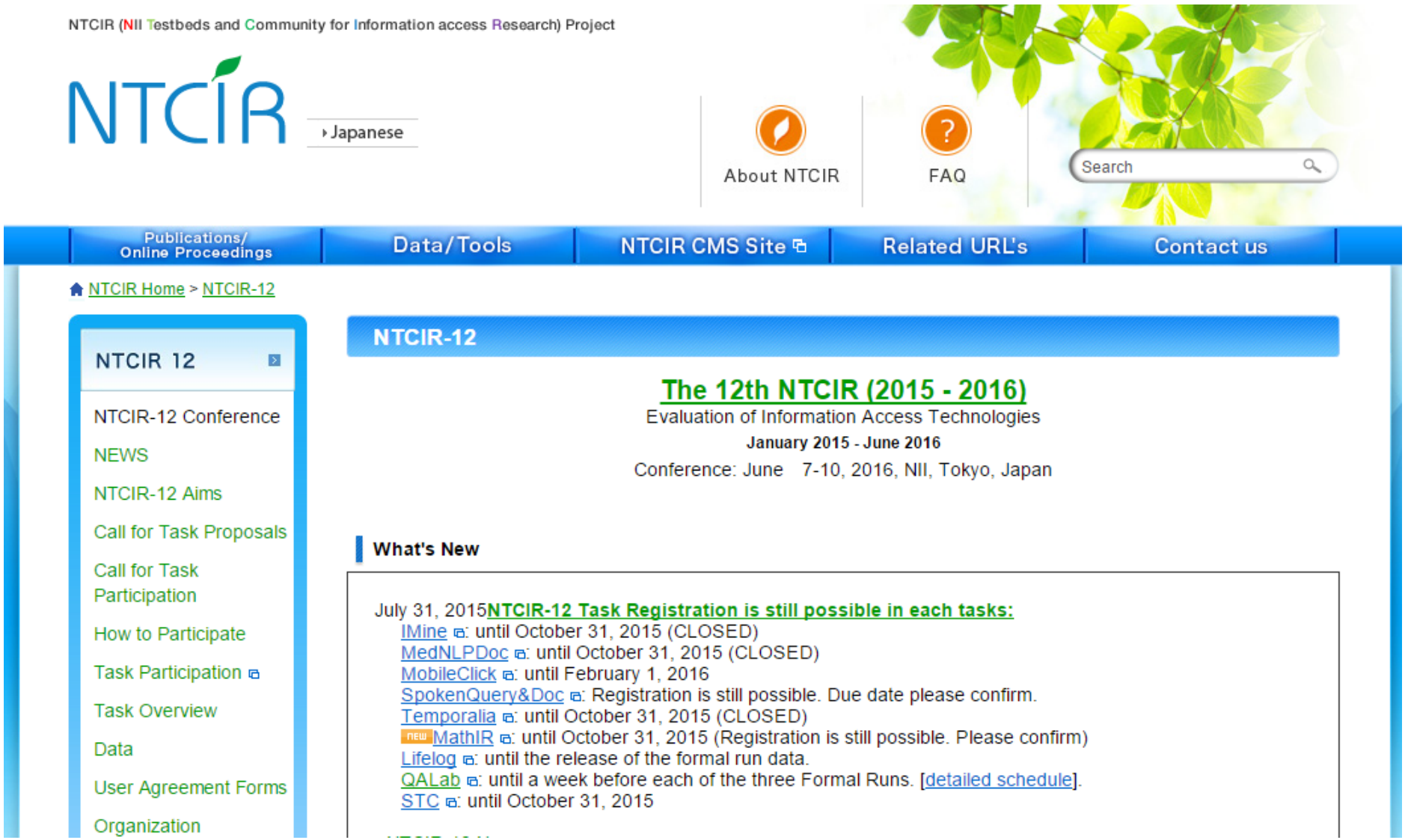
The 12th NTCIR (2015 - 2016)

Evaluation of Information Access Technologies

January 2015 - June 2016

Conference: June 7-10, 2016, NII, Tokyo, Japan

http://research.nii.ac.jp/ntcir/ntcir-12/index.html



http://research.nii.ac.jp/ntcir/ntcir-12/index.html

NTCIR 12 (2015-2016) Tasks

- IMine
- MedNLPDoc
- MobileClick
- SpokenQuery&Doc
- Temporalia
- MathIRNEW
- Lifelog
- QA Lab (QA Lab for Entrance Exam; QALab-2)
- STC

http://research.nii.ac.jp/ntcir/ntcir-12/tasks.html

NTCIR 12 (2015-2016) Schedule

- 31/July302015: Task Registration Due (extended deadline. Registration is still possible in each task. Please see here.)
- 01/July/2015: Document Set Release *
- July-Dec./2015: Dry Run *
- Sep./2015-Feb./2016:Formal Run *
- 01/Feb./2016: Evaluation Results Return
- 01/Feb./2016: Early draft Task Overview Release
- 01/Mar./2016: Draft participant paper submission Due
- 01/May/2016: All camera-ready paper for the Proceedings Due
- 07-10/June/2016:NTCIR-12 Conference & EVIA 2016 in NII, Tokyo, Japan

http://research.nii.ac.jp/ntcir/ntcir-12/dates.html

QA Lab for Entrance Exam (QALab-2)(2015-2016)

- The goal is investigate the real-world complex Question Answering (QA) technologies using Japanese university entrance exams and their English translation on the subject of "World History (世界史)".
- The questions were selected from two different stages - The National Center Test for University Admissions (センター試験, multiple choice-type questions) and from secondary exams at multiple universities (二次試験, complex questions including essays).

http://research.nii.ac.jp/ntcir/ntcir-12/tasks.html

RITE

(Recognizing Inference in Text)

NTCIR-9 RITE (2010-2011)

NTCIR-10 RITE-2 (2012-2013)

NTCIR-11 RITE-VAL (2013-2014)

RITE-2
Recognizing Inference in TExt@NTCIR10

Overview of the Recognizing Inference in TExt (RITE-2) at NTCIR-10

Yotaro Watanabe
Tohoku University

Yusuke Miyao
NII

Junta Mizuno
Tohoku University

Tomohide Shibata
Kyoto University

Hiroshi Kanayama
IBM Research

Cheng-Wei Lee
Academia Sinica

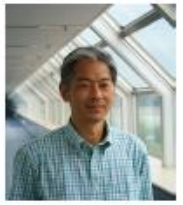
Chuan-Jie Lin
National Taiwan Ocean University

Shuming Shi
MSRA

Teruko Mitamura
CMU

Noriko Kando
NII

Hideki Shima
CMU



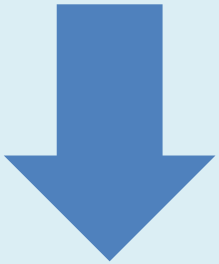
Kohichi Takeda
IBM Research

Source: Yotaro Watanabe, Yusuke Miyao, Junta Mizuno, Tomohide Shibata, Hiroshi Kanayama, Cheng-Wei Lee, Chuan-Jie Lin, Shuming Shi, Teruko Mitamura, Noriko Kando, Hideki Shima and Kohichi Takeda, Overview of the Recognizing Inference in Text (RITE-2) at NTCIR-10, Proceedings of NTCIR-10, 2013,
http://research.nii.ac.jp/ntcir/workshop/OnlineProceedings10/pdf/NTCIR/RITE/01-NTCIR10-RITE2-overview-slides.pdf

Overview of RITE-2

- **RITE-2 is a generic benchmark task that addresses a common semantic inference required in various NLP/IA applications**

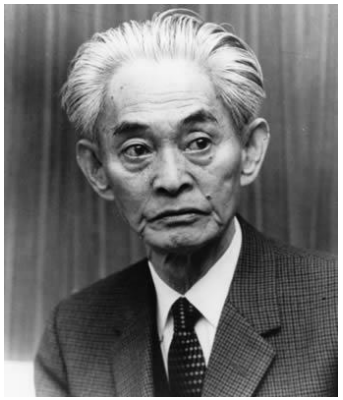
t_1: **Yasunari Kawabata** won the Nobel Prize in Literature for his novel "**Snow Country**."



Can t_2 be inferred from t_1 ? (entailment?)

t_2: **Yasunari Kawabata** is the writer of "**Snow Country**."

Source: Watanabe et al., 2013



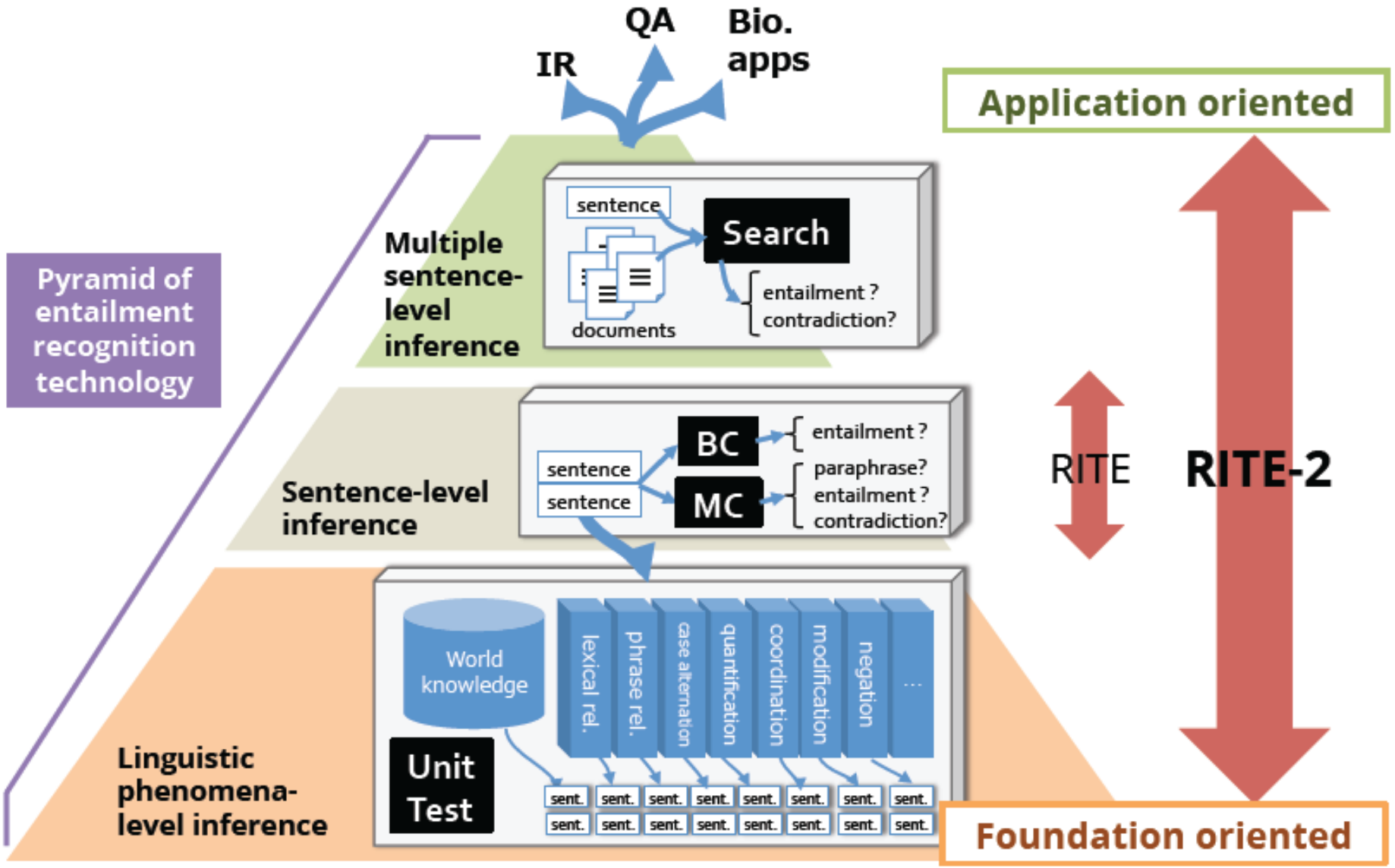
Yasunari Kawabata

Writer

Yasunari Kawabata was a Japanese short story writer and novelist whose spare, lyrical, subtly-shaded prose works won him the Nobel Prize for Literature in 1968, the first Japanese author to receive the award.

http://en.wikipedia.org/wiki/Yasunari_Kawabata

RITE vs. RITE-2



Source: Watanabe et al., 2013

Motivation of RITE-2

- Natural Language Processing (NLP) / Information Access (IA) applications
  - Question Answering, Information Retrieval, Information Extraction, Text Summarization, Automatic evaluation for Machine Translation, Complex Question Answering
- The current entailment recognition systems have not been mature enough
  - The highest accuracy on Japanese BC subtask in NTCIR-9 RITE was only 58%
  - There is still enough room to address the task to advance entailment recognition technologies

Source: Watanabe et al., 2013

BC and MC subtasks in RITE-2

t_1: **Yasunari Kawabata** won the Nobel Prize in Literature for his novel "**Snow Country**."

t_2: **Yasunari Kawabata** is the writer of "**Snow Country**."

- **BC subtask**

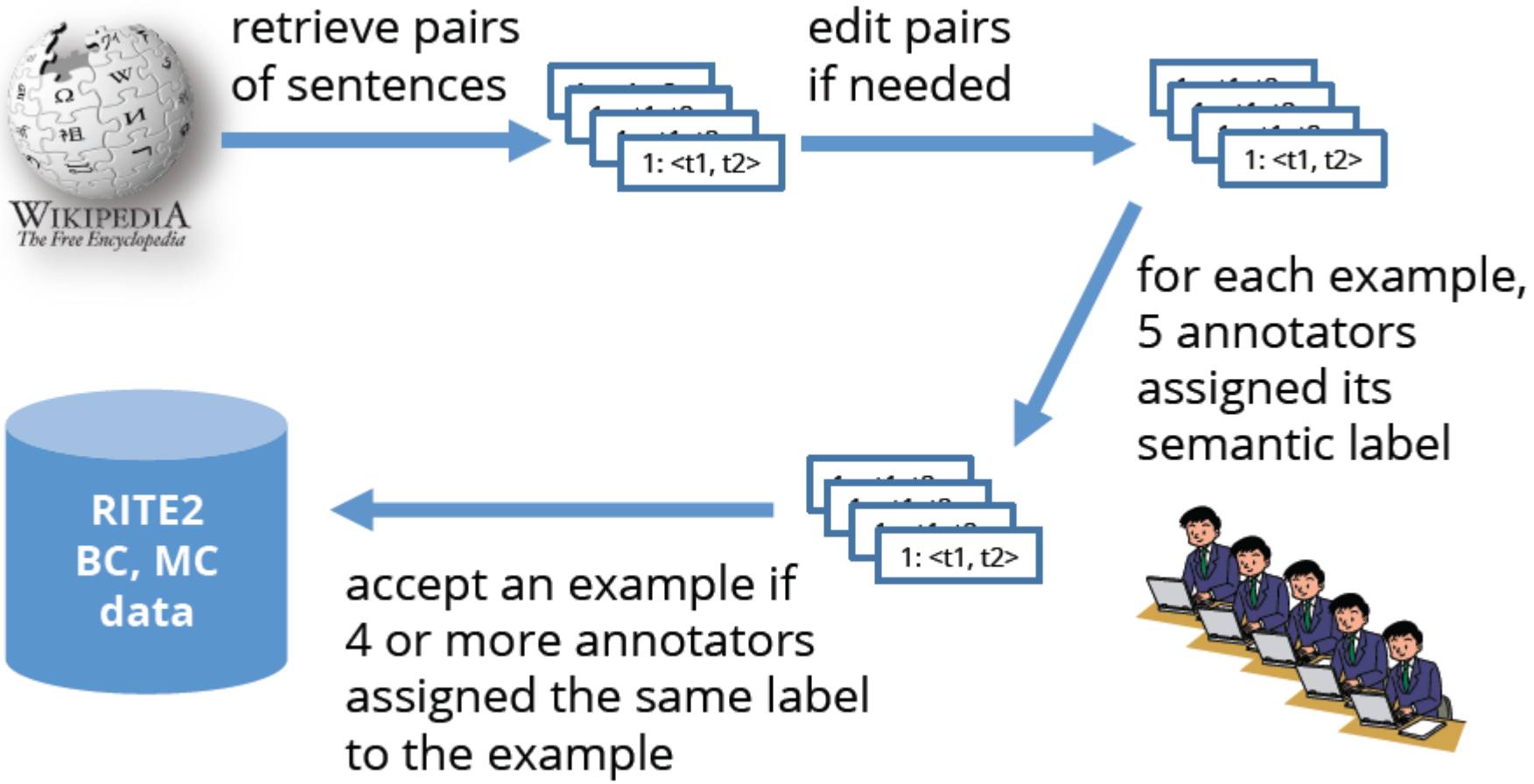
  - Entailment (t_1 entails t_2) or Non-Entailment (otherwise)

- **MC subtask**

  - Bi-directional Entailment (t_1 entails t_2 & t_2 entails t_1)
  - Forward Entailment (t_1 entails t_2 & t_2 does not entail t_1)
  - Contradiction (t_1 contradicts t_2 or cannot be true at the same time)
  - Independence (otherwise)

Source: Watanabe et al., 2013

Development of BC and MC data

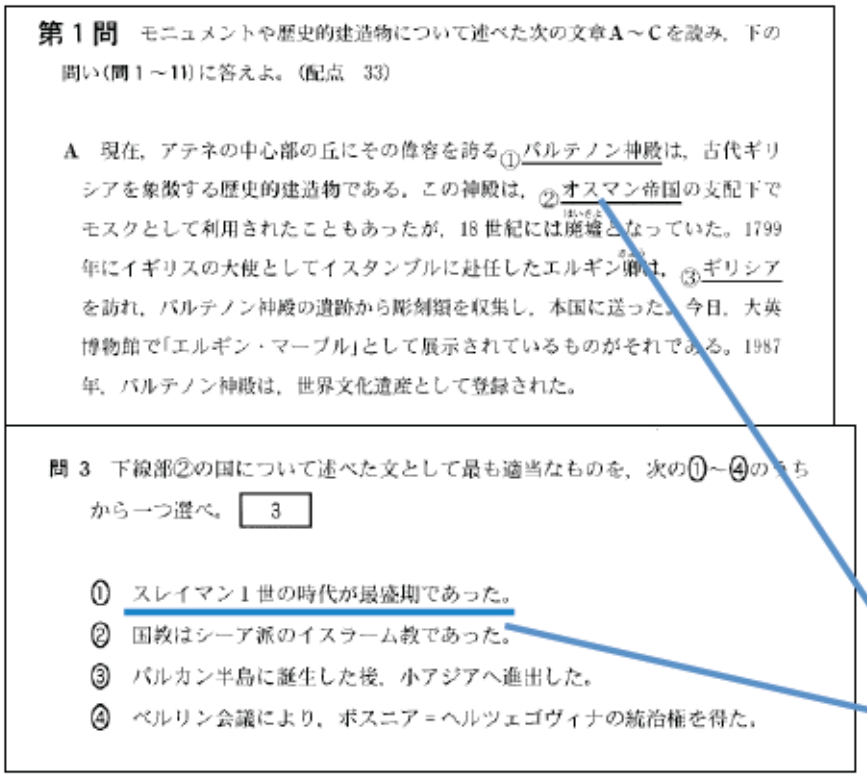


Source: Watanabe et al., 2013

Entrance Exam subtasks (Japanese only)

Entrance exam problem

National Center Test for University Admission
(*Daigaku Nyushi Center Shiken*)

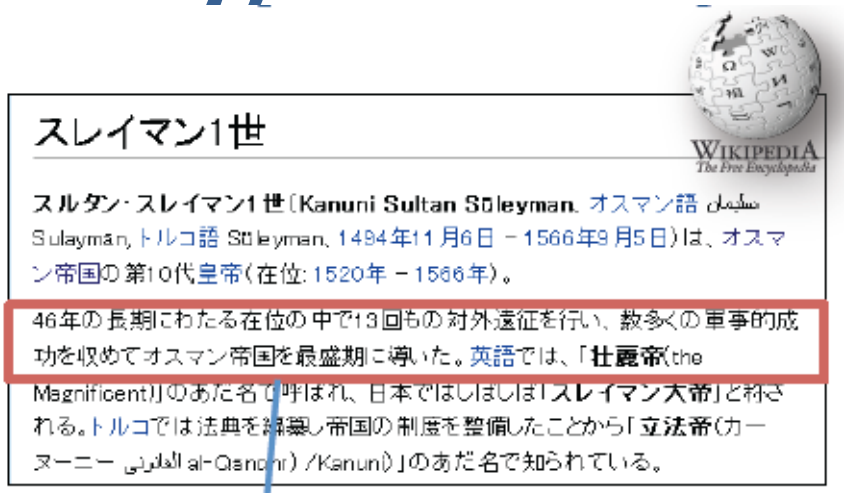


第１問　モニュメントや歴史的建造物について述べた次の文章Ａ～Ｃを読み，下の問い（問１～11）に答えよ。（配点　33）

Ａ　現在，アテネの中心部の丘にその偉容を誇る①パルテノン神殿は，古代ギリシアを象徴する歴史的建造物である。この神殿は，②オスマン帝国の支配下でモスクとして利用されたこともあったが，18世紀には廃墟となっていた。1799年にイギリスの大使としてイスタンブルに赴任したエルギン卿は，③ギリシアを訪れ，パルテノン神殿の遺跡から彫刻類を収集し，本国に送った。今日，大英博物館で「エルギン・マーブル」として展示されているものがそれである。1987年，パルテノン神殿は，世界文化遺産として登録された。

問３　下線部②の国について述べた文として最も適当なものを，次の①～④のうちから一つ選べ。　3

① スレイマン１世の時代が最盛期であった。
② 国教はシーア派のイスラーム教であった。
③ バルカン半島に誕生した後，小アジアへ進出した。
④ ベルリン会議により，ボスニア＝ヘルツェゴヴィナの統治権を得た。



スレイマン1世

スルタン・スレイマン1世（Kanuni Sultan Süleyman, オスマン語 سليمان Sulayman, トルコ語 Süleyman, 1494年11月6日 – 1566年9月5日）は、オスマン帝国の第10代皇帝（在位: 1520年 – 1566年）。

46年の長期にわたる在位の中で13回もの対外遠征を行い、数多くの軍事的成功を収めてオスマン帝国を最盛期に導いた。英語では、「壮麗帝（the Magnificent）」のあだ名で呼ばれ、日本ではしばしば「スレイマン大帝」と称される。トルコでは法典を編纂し帝国の制度を整備したことから「立法者（カーヌーニー القانوني al-Qanuni）/Kanuni）」のあだ名で知られている。

t_1: スレイマン1世は数多くの軍事的成功を収めてオスマン帝国を最盛期に導いた．(Suleiman I contributed in a lot of military successes and led the Ottoman Empire to its peak.

t_2: オスマン帝国ではスレイマン1世の時代が最盛期であった．(The Ottoman Empire's peak was during the reign of Suleiman I).

Source: Watanabe et al., 2013

Entrance Exam subtask: BC and Search

- **Entrance Exam BC**
  - Binary-classification problem (Entailment or Nonentailment)
  - t1 and t2 are given
- **Entrance Exam Search**
  - Binary-classification problem (Entailment or Nonentailment)
  - t2 and a set of documents are given
    - Systems are required to search sentences in Wikipedia and textbooks to decide semantic labels

Source: Watanabe et al., 2013

UnitTest (Japanese only)

- **Motivation**
  - Evaluate how systems can handle linguistic
  - phenomena that affects entailment relations
- **Task definition**
  - Binary classification problem (same as BC subtask)

Source: Watanabe et al., 2013

RITE4QA (Chinese only)

- **Motivation**
  - Can an entailment recognition system rank a set of unordered answer candidates in QA?
- **Dataset**
  - Developed from NTCIR-7 and NTCIR-8 CLQA data
    - t1: answer-candidate-bearing sentence
    - t2: a question in an affirmative form
- **Requirements**
  - Generate confidence scores for ranking process

Source: Watanabe et al., 2013

Evaluation Metrics

- **Macro F1 and Accuracy (BC, MC, ExamBC, ExamSearch and UnitTest)**

$$MacroF1 = \frac{1}{|C|}\sum_{c \in C} F1_c \qquad Accuracy = 100 \times \frac{N_{correct}}{N_{examples}}$$

- **Correct Answer Ratio (Entrance Exam)**
  - Y/N labels are mapped into selections of answers and calculate accuracy of the answers
- **Top1 and MRR (RITE4QA)**

$$Top1 = \frac{1}{|Q|}\sum_{i=1}^{|Q|}[\text{top answer is correct}] \qquad MRR = \frac{1}{|Q|}\sum_{i=1}^{|Q|}\frac{1}{rank_i}$$

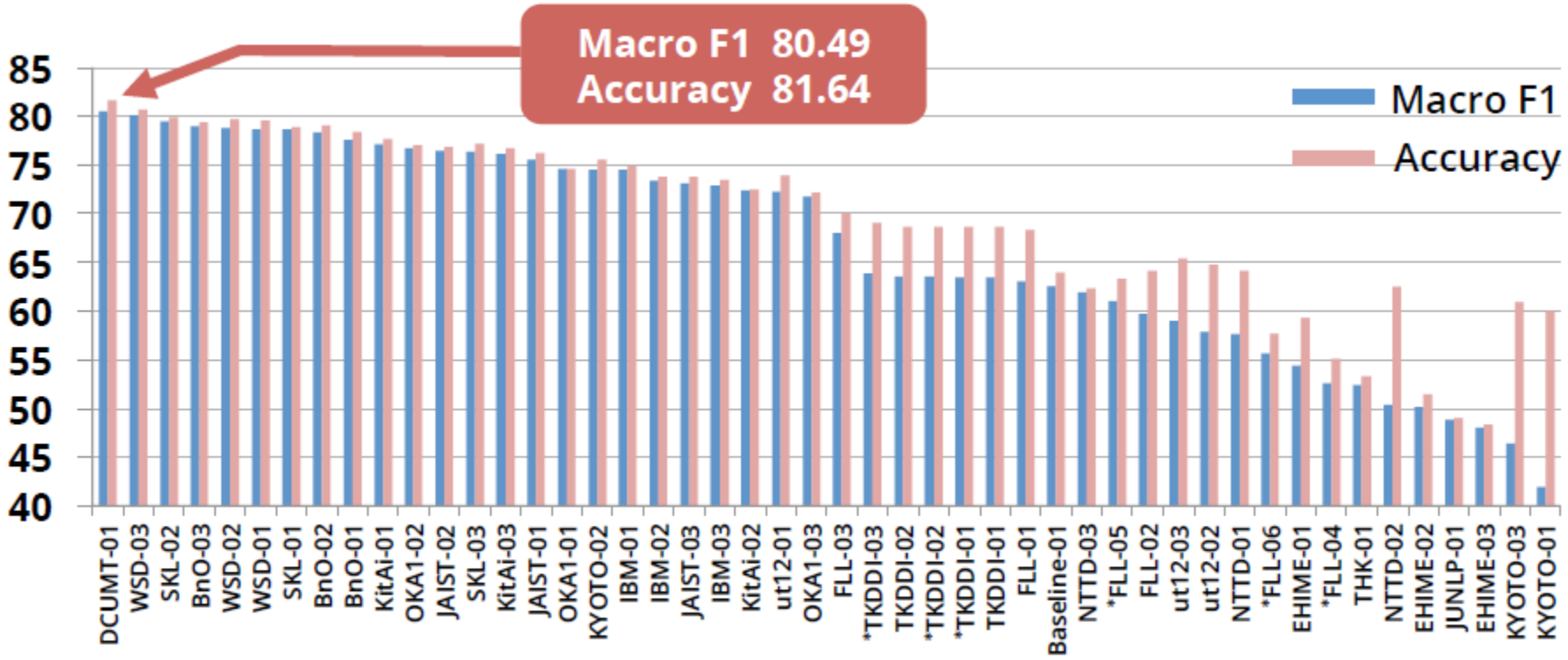
Source: Watanabe et al., 2013

RITE-2
Recognizing Inference in TExt@NTCIR10

Countries/Regions of Participants

Source: Watanabe et al., 2013

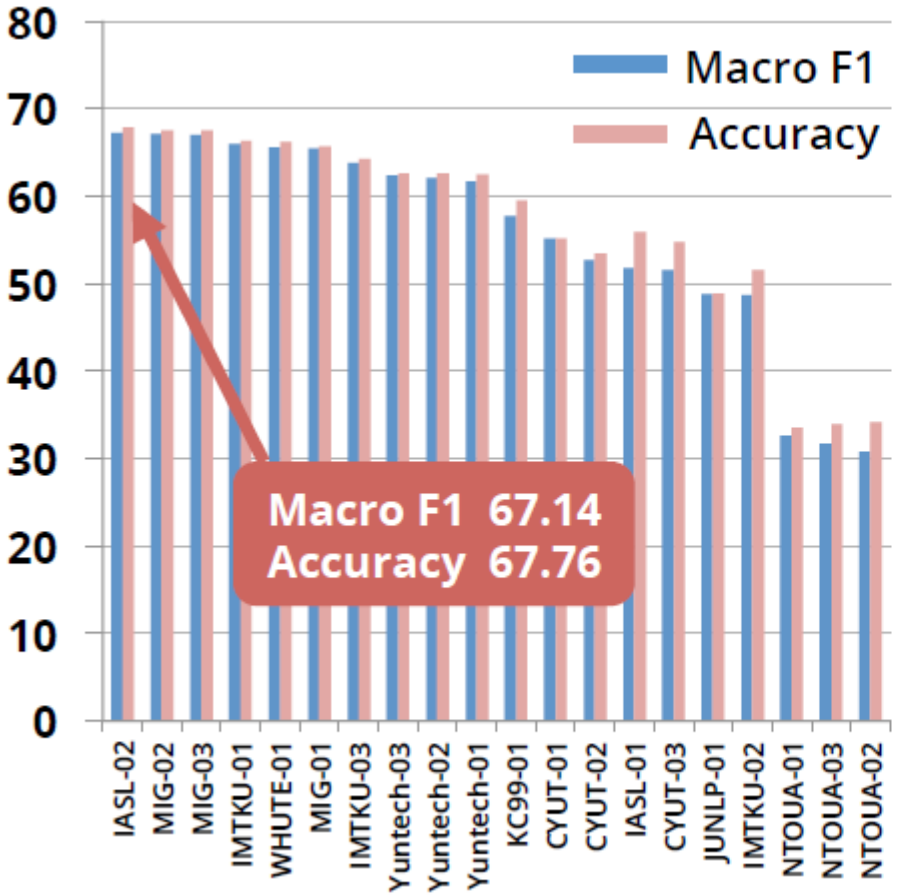
Formal Run Results: BC (Japanese)

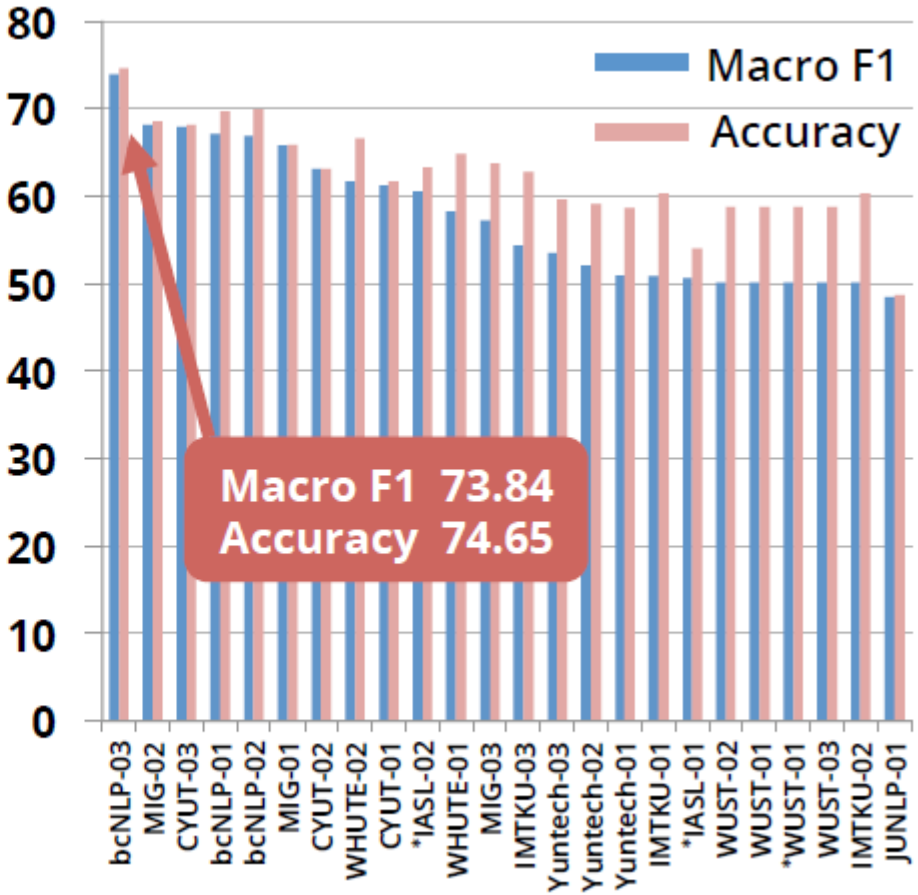


- **The best system achieved over 80% of accuracy (The highest score in BC subtask at RITE was 58%)**
- **The difference is caused by**
  - Advancement of entailment recognition technologies
  - Strict data filtering in the data development

Source: Watanabe et al., 2013

BC (Traditional/Simplified Chinese)

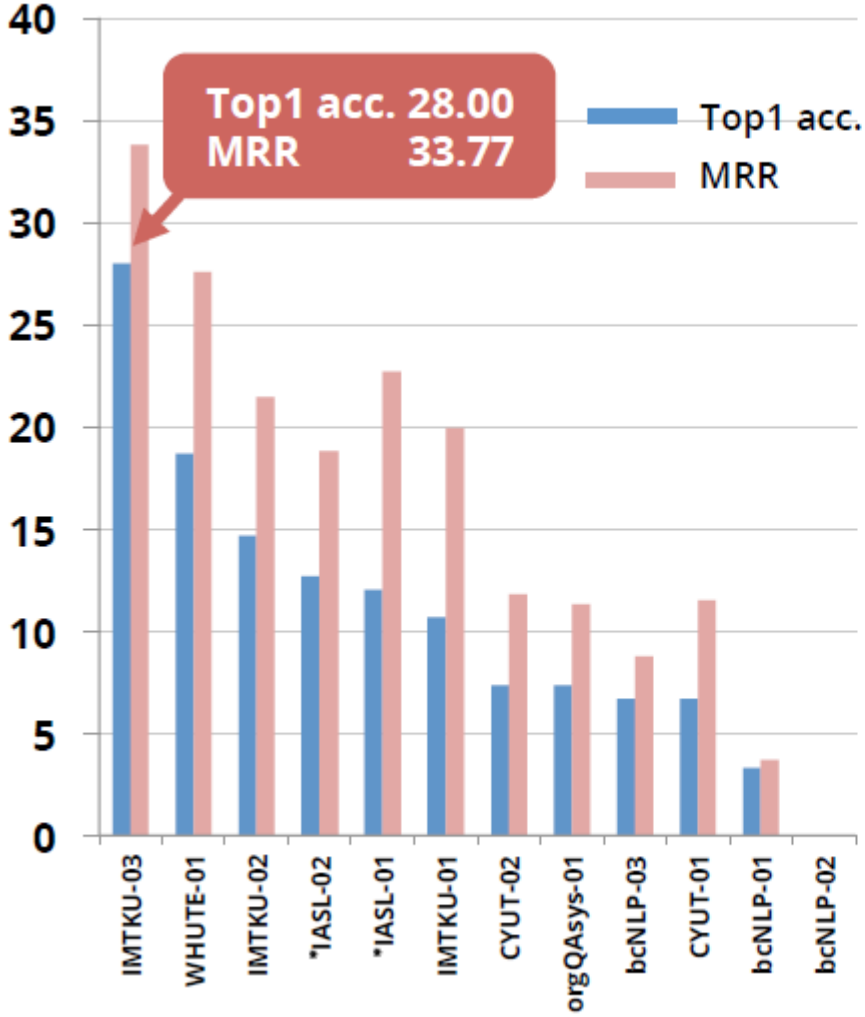


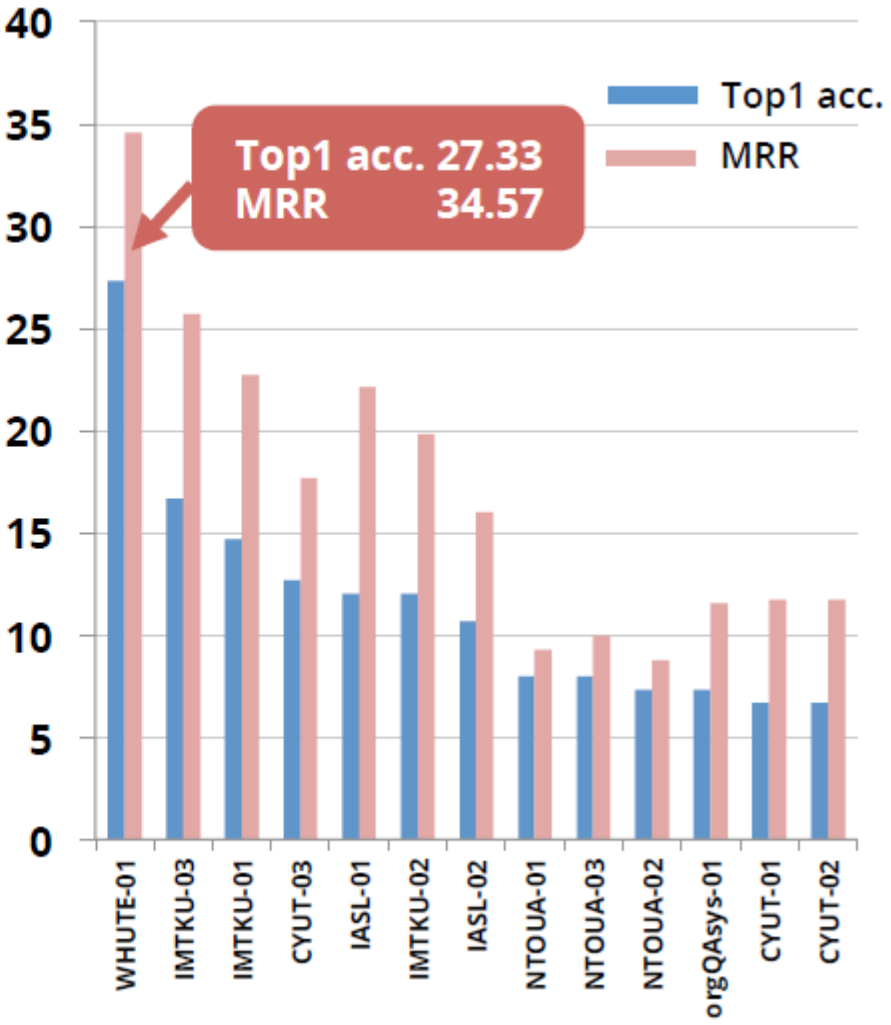


The top scores are almost the same as those in NTCIR-9 RITE

Source: Watanabe et al., 2013

RITE4QA (Traditional/Simplified Chinese)





Source: Watanabe et al., 2013

Participant's approaches in RITE-2

- **Category**
  - Statistical (50%)
  - Hybrid (27%)
  - Rule-based (23%)
- **Fundamental approach**
  - Overlap-based (77%)
  - Alignment-based (63%)
  - Transformation-based (23%)

Source: Watanabe et al., 2013

Summary of types of information explored in RITE-2

- Character/word overlap (85%)
- Syntactic information (67%)
- Temporal/numerical information (63%)
- Named entity information (56%)
- Predicate-argument structure (44%)
- Entailment relations (30%)
- Polarity information (7%)
- Modality information (4%)

Source: Watanabe et al., 2013

Summary of Resources Explored in RITE-2

- **Japanese**
  - Wikipedia (10)
  - Japanese WordNet (9)
  - ALAGIN Entailment DB (5)
  - Nihongo Goi-Taikei (2)
  - Bunruigoihyo (2)
  - Iwanami Dictionary (2)
- **Chinese**
  - Chinese WordNet (3)
  - TongYiCi CiLin (3)
  - HowNet (2)

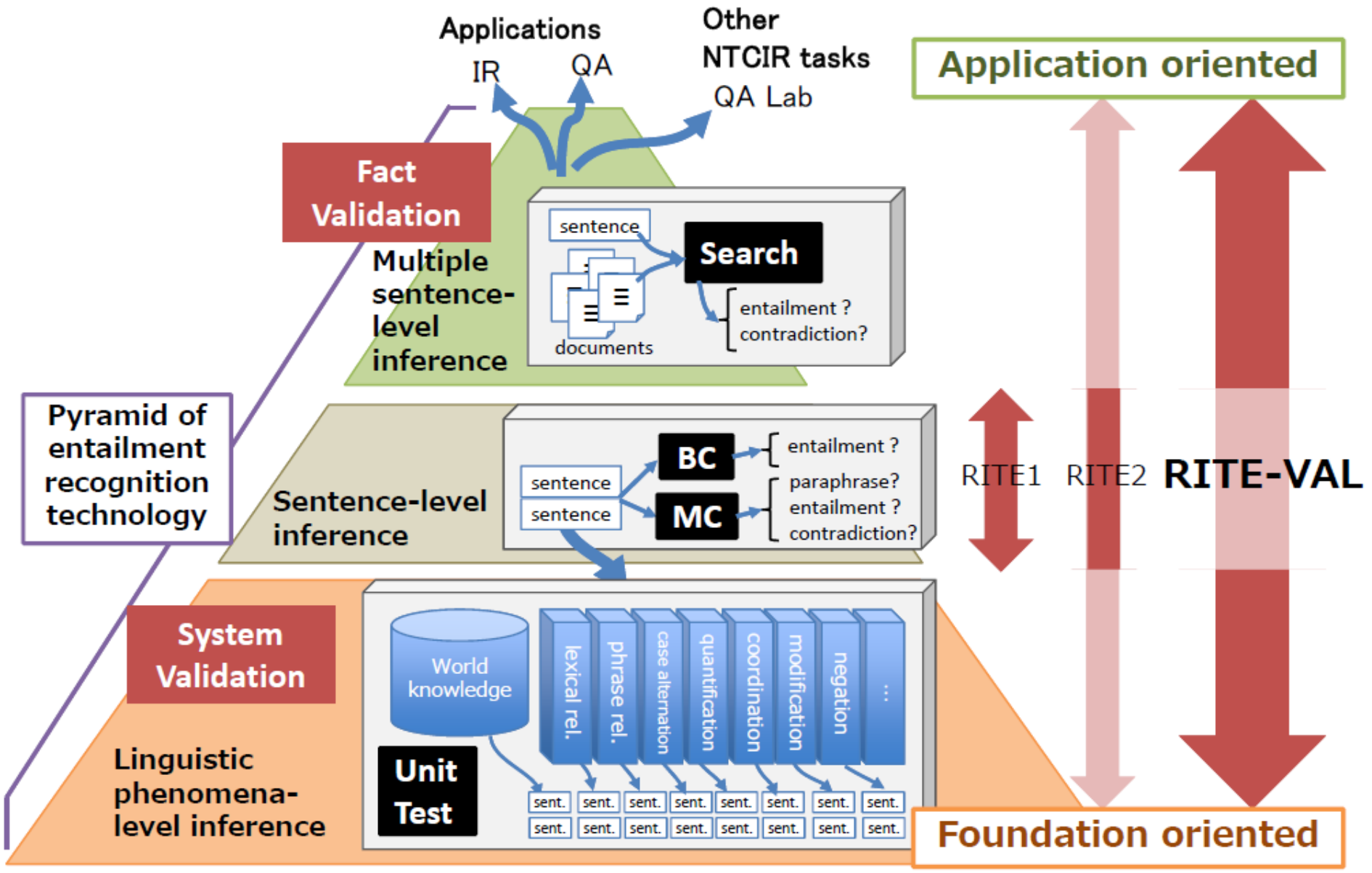
Source: Watanabe et al., 2013

Advanced approaches in RITE-2

- **Logical approaches**
  - Dependency-based Compositional Semantics (DCS) [BnO], Markov Logic [EHIME], Natural Logic [THK]
- **Alignment**
  - GIZA [CYUT], ILP [FLL], Labeled Alignment [bcNLP, THK]
- **Search Engine**
  - Google and Yahoo [DCUMT]
- **Deep Learning**
  - RNN language models [DCUMT]
- **Probabilistic Models**
  - N-gram HMM [DCUMT], LDA [FLL]
- **Machine Translation**
  - [JUNLP, JAIST, KC99]

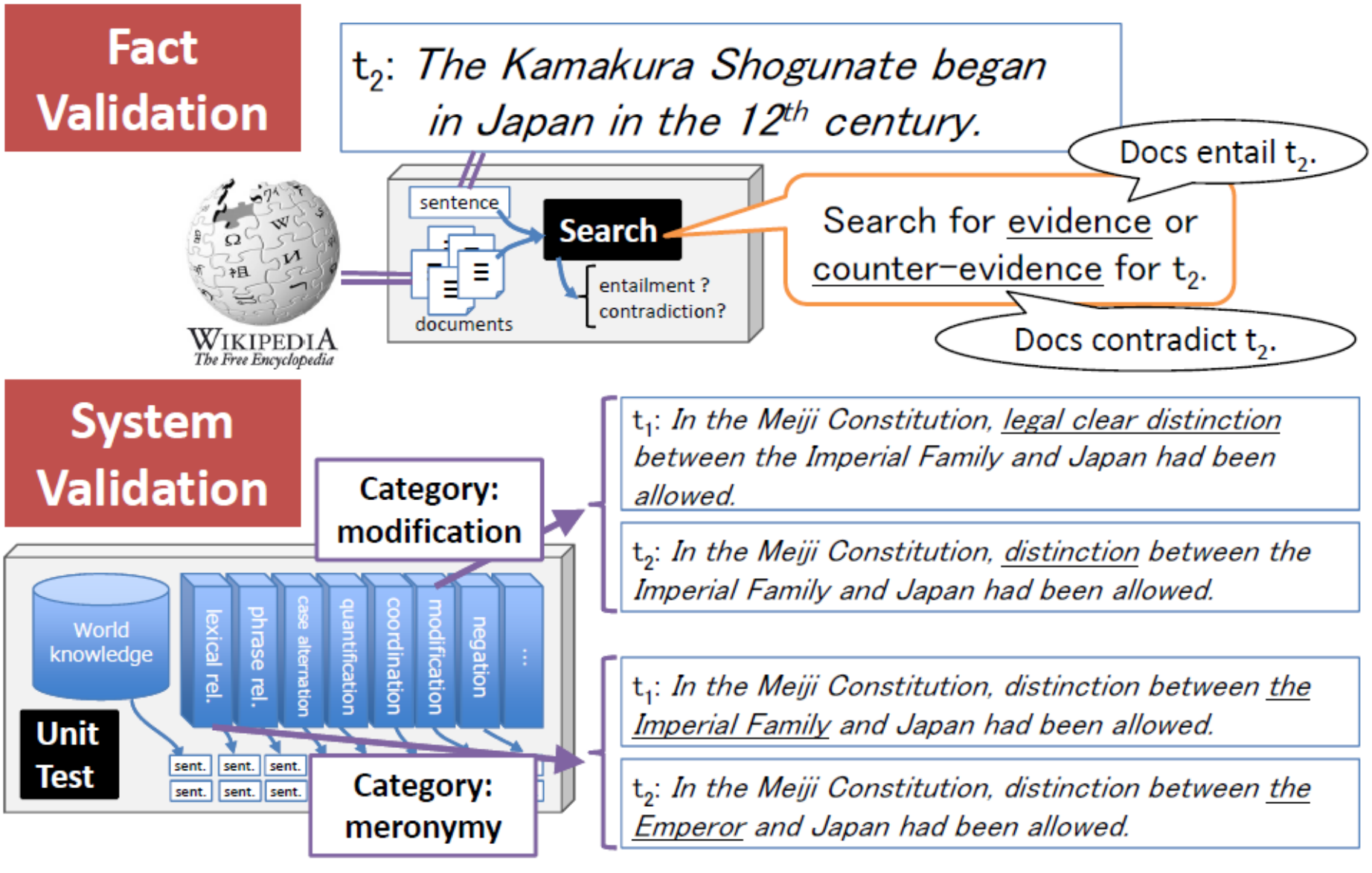
Source: Watanabe et al., 2013

RITE-VAL



Source: Matsuyoshi et al., 2013

Main two tasks of RITE-VAL



Source: Matsuyoshi et al., 2013

Tamkang University
淡江大學

2011

Tamkang University

IMTKU Textual Entailment System for Recognizing Inference in Text at NTCIR-9 RITE

Department of Information Management
Tamkang University, Taiwan

Min-Yuh Day **Chun Tu**

myday@mail.tku.edu.tw

NTCIR-9 Workshop, December 6-9, 2011, Tokyo, Japan

Tamkang University
淡江大學

2013

Tamkang University

IMTKU Textual Entailment System for Recognizing Inference in Text at NTCIR-10 RITE-2

Department of Information Management
Tamkang University, Taiwan


Min-Yuh Day


Chun Tu


Hou-Cheng Vong


Shih-Wei Wu


Shih-Jhen Huang

myday@mail.tku.edu.tw

NTCIR-10 Conference, June 18-21, 2013, Tokyo, Japan

IMTKU Textual Entailment System for Recognizing Inference in Text at NTCIR-11 RITE-VAL 2014

Tamkang University

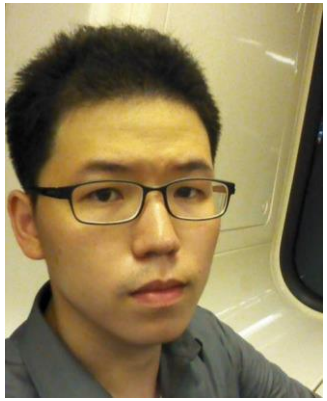
淡江大學

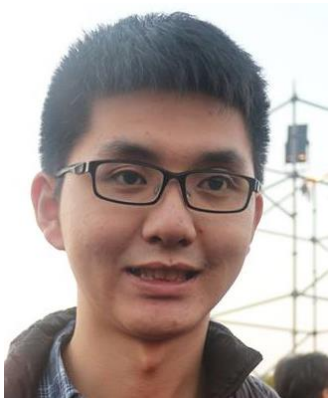

Min-Yuh Day


Ya-Jung Wang


Che-Wei Hsu

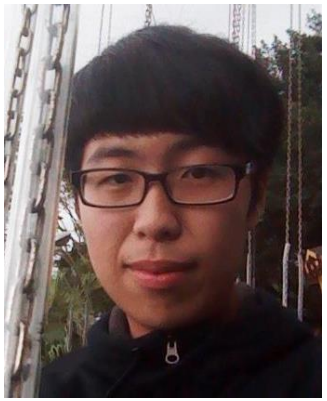

En-Chun Tu


Huai-Wen Hsu


Yu-An Lin


Shang-Yu Wu


Yu-Hsuan Tai


Cheng-Chia Tsai

NTCIR-11 Conference, December 8-12, 2014, Tokyo, Japan

IMTKU Question Answering System for Entrance Exam at NTCIR-12 QALab-2

Tamkang University

淡江大學

2016


Min-Yuh Day


Cheng-Chia Tsai


Wei-Chun Chung


Hsiu-Yuan Chang


Lin-Jin Kun


Yuan-Jie Tsai


Tzu-Jui Sun


Yu-Ming Guo


Yue-Da Lin


Wei-Ming Chen


Yun-Da Tsai

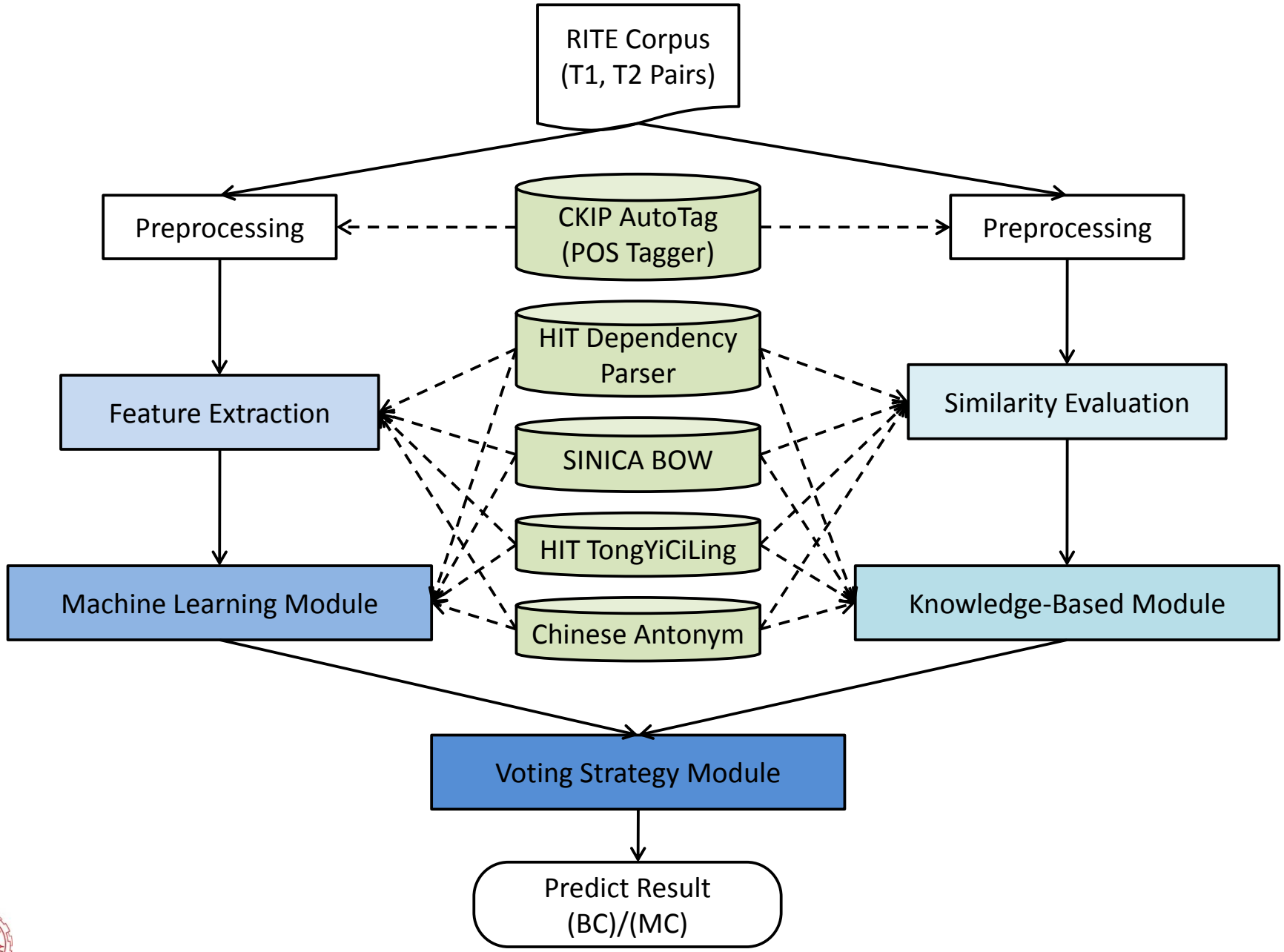

Cheng-Jhih Han


Yi-Jing Lin

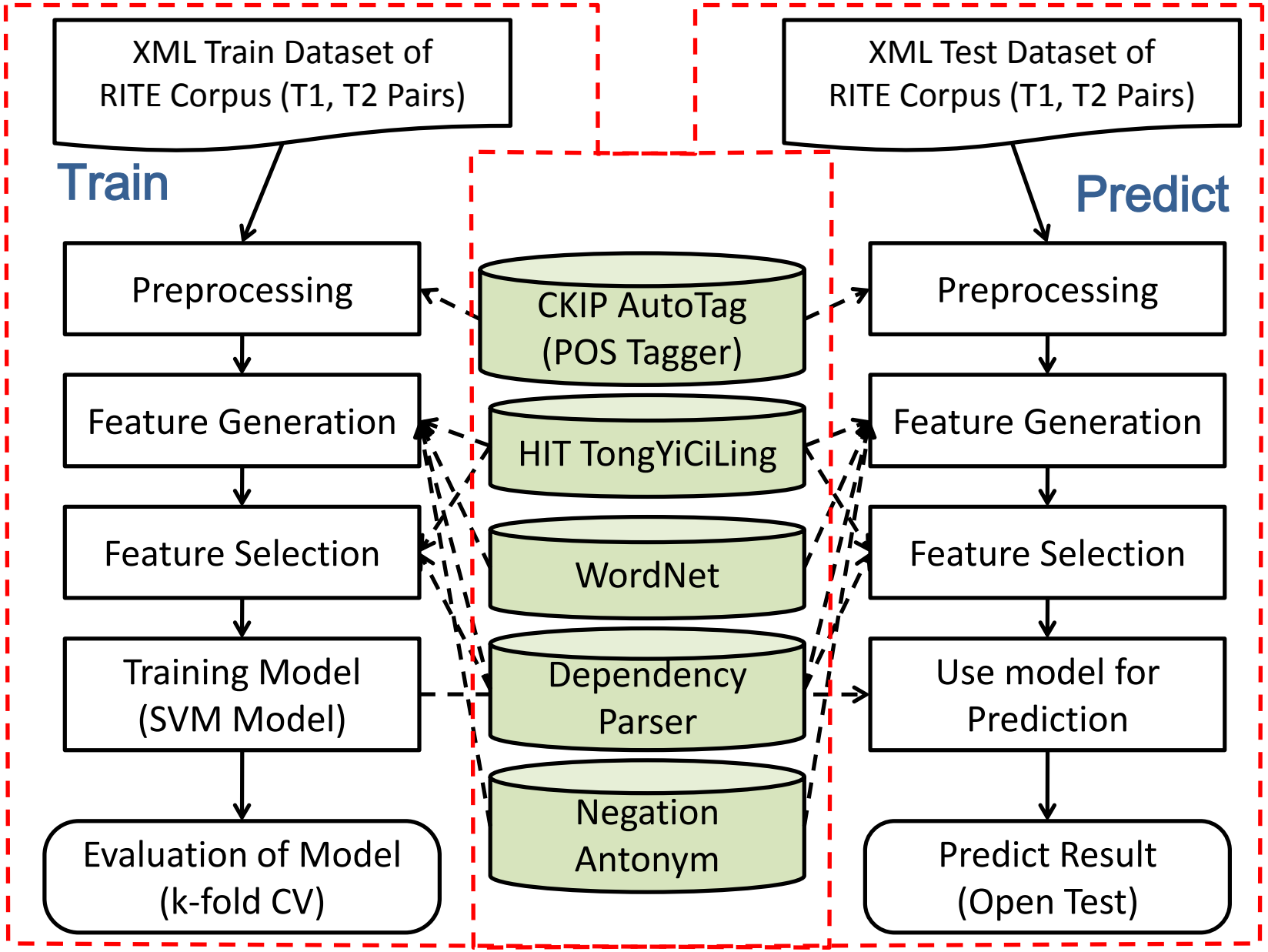

Yi-Heng Chiang


Ching-Yuan Chien

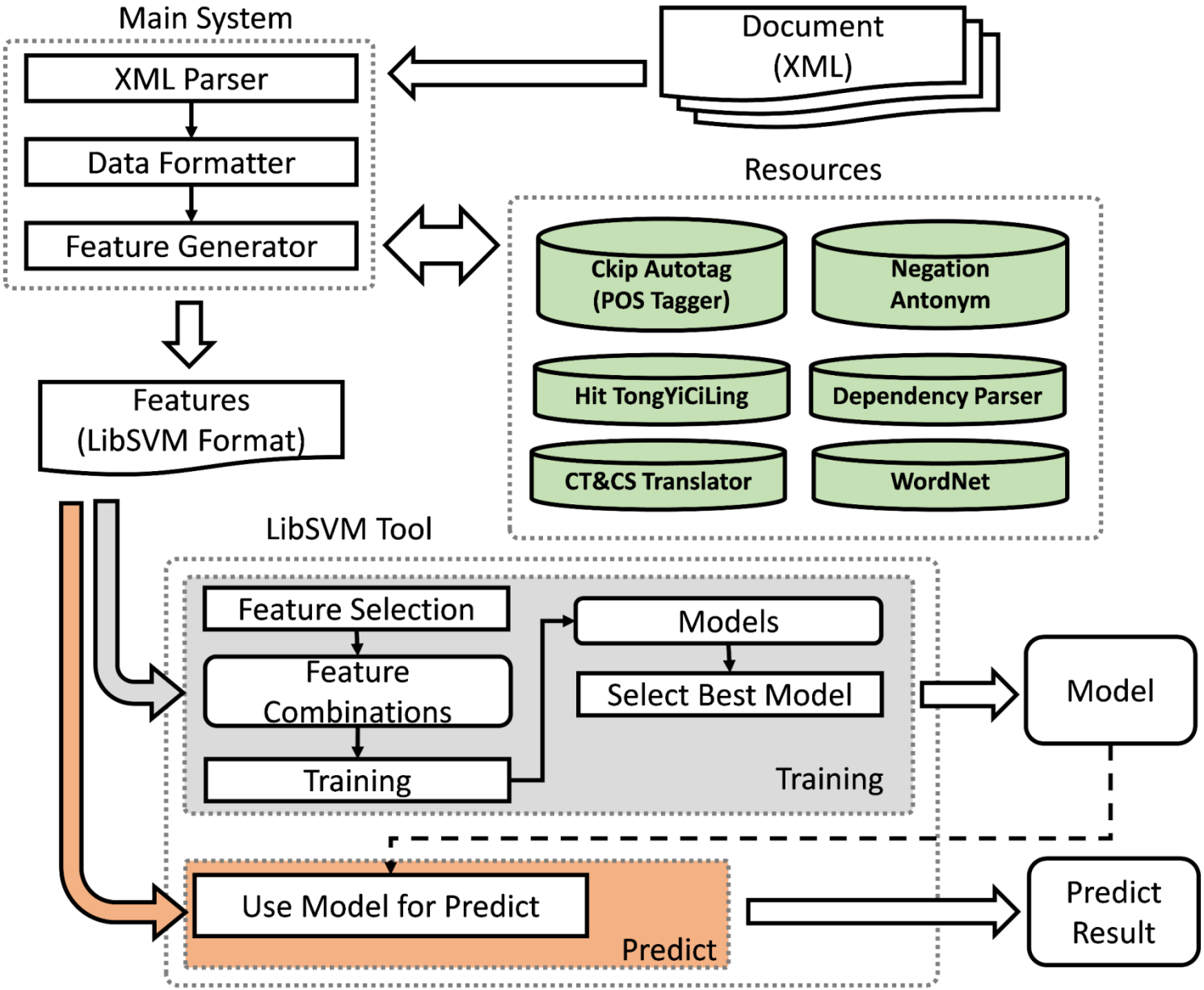
IMTKU System Architecture for NTCIR-9 RITE



IMTKU System Architecture for NTCIR-10 RITE-2



IMTKU System Framework for NTCIR -11 RITE-VAL



IMTKU at NTCIR

- The first place in the CS-RITE4QA subtask of the NTCIR-10 Recognizing Inference in TExt (RITE) task. (2013)
- The second place in the CT-RITE4QA subtask of the NTCIR-10 Recognizing Inference in TExt (RITE) task. (2013)
- The first place in the CT-RITE4QA subtask of the NTCIR-9 Recognizing Inference in TExt (RITE) task. (2011)
- The first place in the CS-RITE4QA subtask of the NTCIR-9 Recognizing Inference in TExt (RITE) task. (2011)
- The second place in the CT-MC subtask of the NTCIR-9 Recognizing Inference in TExt (RITE) task. (2011)

Summary

- Big Data Analytics on Social Media
- Analyzing the Social Web:
  Social Network Analysis
- NTCIR 12 QALab-2 Task

References

- Jiawei Han and Micheline Kamber (2011),
  Data Mining: Concepts and Techniques, Third Edition, Elsevier
- Jennifer Golbeck (2013),
  Analyzing the Social Web, Morgan Kaufmann
- Stephan Kudyba (2014),
  Big Data, Mining, and Analytics: Components of Strategic Decision Making, Auerbach Publications
- Hiroshi Ishikawa (2015),
  Social Big Data Mining, CRC Press

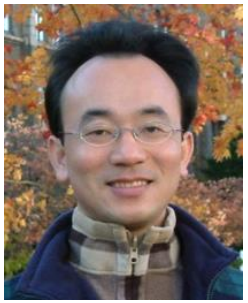
Tamkang University
淡江大學

Tamkang University

Q & A

Big Data Analytics on Social Media (社群媒體大數據分析)

Time: 2015/12/25 (14:00-15:30)
Place: S402, Ming Chuan University



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/

2015-12-25