

# 大數據行銷研究

## Big Data Marketing Research



Tamkang  
University  
淡江大學

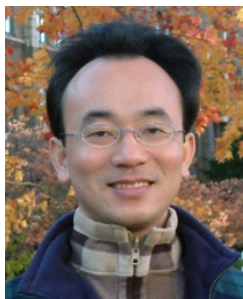
# 社群運算與大數據分析

## (Social Computing and Big Data Analytics)

1051BDMR08

MIS EMBA (M2262) (8638)

Thu, 12,13,14 (19:20-22:10) (D409)



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/>

2016-11-25



# 課程大綱 (Syllabus)

週次 (Week)	日期 (Date)	內容 (Subject/Topics)
1	2016/09/16	中秋節 (調整放假一天) (Mid-Autumn Festival Holiday)(Day off)
2	2016/09/23	大數據行銷研究課程介紹 (Course Orientation for Big Data Marketing Research)
3	2016/09/30	資料科學與大數據行銷 (Data Science and Big Data Marketing)
4	2016/10/07	大數據行銷分析與研究 (Big Data Marketing Analytics and Research)
5	2016/10/14	測量構念 (Measuring the Construct)
6	2016/10/21	測量與量表 (Measurement and Scaling)

# 課程大綱 (Syllabus)

週次 (Week)	日期 (Date)	內容 (Subject/Topics)
7	2016/10/28	大數據行銷個案分析 I (Case Study on Big Data Marketing I)
8	2016/11/04	探索性因素分析 (Exploratory Factor Analysis)
9	2016/11/11	確認性因素分析 (Confirmatory Factor Analysis)
10	2016/11/18	期中報告 (Midterm Presentation)
11	2016/11/25	社群運算與大數據分析 (Social Computing and Big Data Analytics)
12	2016/12/02	社會網路分析 (Social Network Analysis)

# 課程大綱 (Syllabus)

週次 (Week)	日期 (Date)	內容 (Subject/Topics)
13	2016/12/09	大數據行銷個案分析 II (Case Study on Big Data Marketing II)
14	2016/12/16	社會網絡分析量測與實務 (Measurements and Practices of Social Network Analysis)
15	2016/12/23	大數據情感分析 (Big Data Sentiment Analysis)
16	2016/12/30	金融科技行銷研究 (FinTech Marketing Research)
17	2017/01/06	期末報告 I (Term Project Presentation I)
18	2017/01/13	期末報告 II (Term Project Presentation II)

# Outline

- Social Computing
- Big Data Analysis

# 教育部資通人才培育計畫

## 社群運算與巨量資料

### 課程四大模組

- (1) 「社群媒體」 (Social Media)  
(政治大學)
- (2) 「資料科學」 (Data Science)  
(政治大學)
- (3) 「分析技術」 (Analytics Technology)  
(高雄大學) (淡江大學)
- (4) 「領域應用」 (Domain Application)  
(淡江大學) (政治大學)

# 1. 「社群媒體」(Social Media) (政治大學)

- 探討社群媒體和資料分析的概念，以個案方式教學

## 2. 「資料科學」 (Data Science) (政治大學)

- 探討 Data Thinking 和 EDA 等，  
與DSP或痞客邦合作



### 3. 「分析技術」(Analytics Technology) (高雄大學)(淡江大學)

- 列舉重要的分析方法，包括社會網絡分析，文字探勘分析技術簡介。
  - \* 社會網絡分析(高雄大學)
  - \* 社會網絡量測(高雄大學)
  - \* 社會網絡分析工具(高雄大學)
  - \* 文字探勘分析技術簡介(淡江大學)

## 4. 「領域應用」 (Domain Application) (淡江大學) (政治大學)

- 區分 Domain Knowledge ， 聚焦探討各種商業行銷和輿情分析等
  - \* 社群媒體行銷分析 (淡江大學)
  - \* 社群媒體情感分析 (淡江大學)

# Social Computing

# **Social Network Analysis**

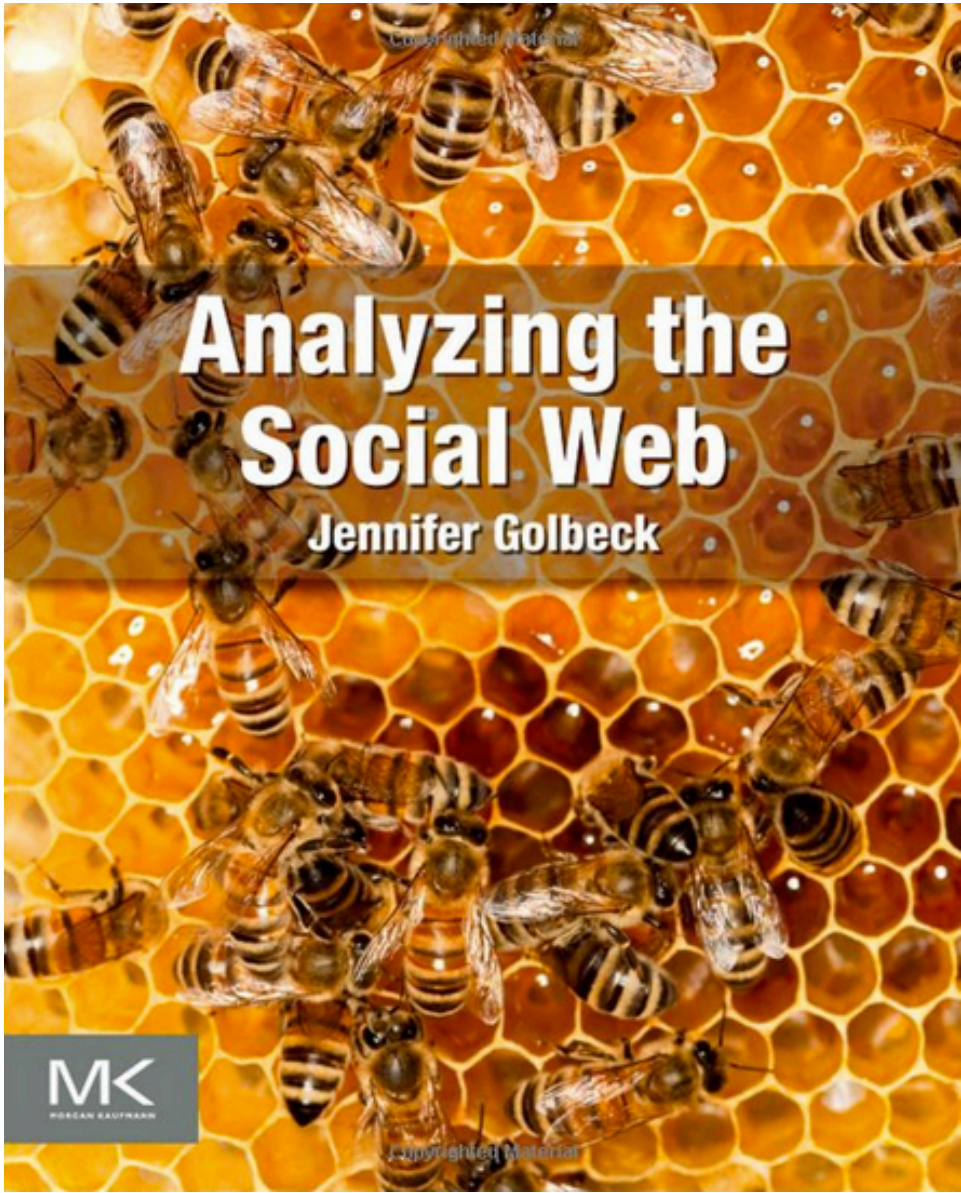
# Social Computing

- Social Network Analysis
- Link mining
- Community Detection
- Social Recommendation

# Business Insights with Social Analytics

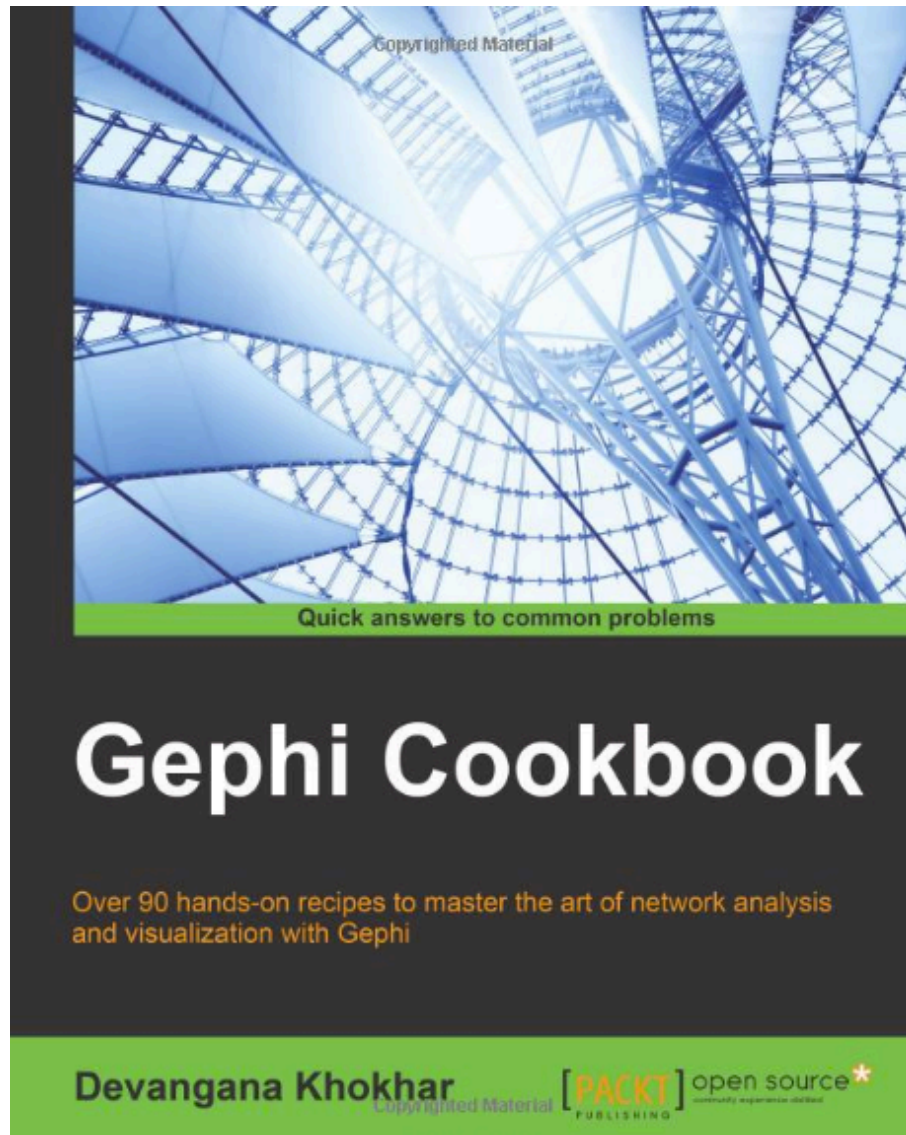
# Analyzing the Social Web: Social Network Analysis

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





# Devangana Khokhar (2015), Gephi Cookbook, Packt Publishing



# Social Network Analysis (SNA)

## Facebook TouchGraph

TouchGraph Photos x

box.touchgraph.com/facebook/TGFacebookBrowser.php?&signed\_request=Gi-L3\_6HrZ0S3SjxAXGdHR0rhMzqBjUnvFJ9vE4W6vg.eyJhbGdvcm00aG0iOiJITUFDI☆

Profiles Networks

Show Top 100 Friends Show All Friends Upload Advanced Restart

Zoom: Spacing:

Min-Yuh Day  
 Networks: None  
 Mutual Friends: 681

Facebook Profile

Network All All List Photo

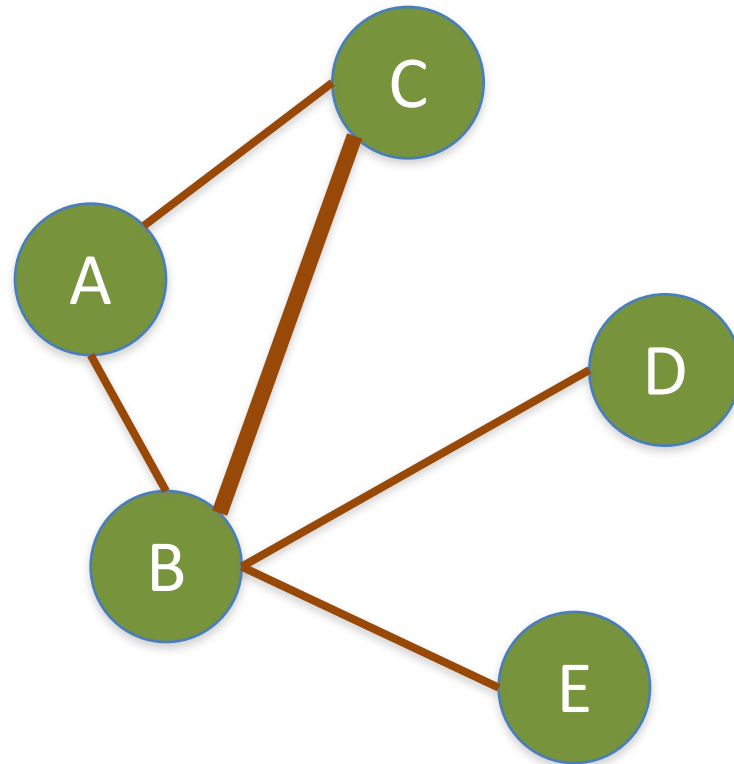
Name	Rank #	Friend #	民
Min-Yuh Day	1	681	
Gladys Hsieh	2	85	
黃西田	3	74	
施盛賓	4	67	
John Lee	5	104	
Kevin Tu	6	61	
Yung Yu Shih	7	45	
Wei Chen	8	107	
Chichang Jou	9	50	
Allen Green	10	81	
黃煒勳	11	65	
梁德昭	12	44	
Eric Chen	13	51	
吳錦波	14	39	
Jessica Tien	15	49	
蔡名宜	16	112	
Enrico Lu	17	59	
YaHan Hsieh	18	64	
王慧雯	19	56	
薛聖譚	20	80	
蝦米	21	73	

ICCU

powered by TouchGraph

# Graph Theory

# Graph



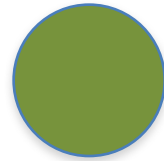
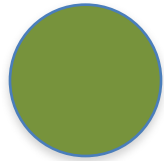
# Graph

$$g = (V, E)$$

# Vertex (Node)



# Vertices (Nodes)

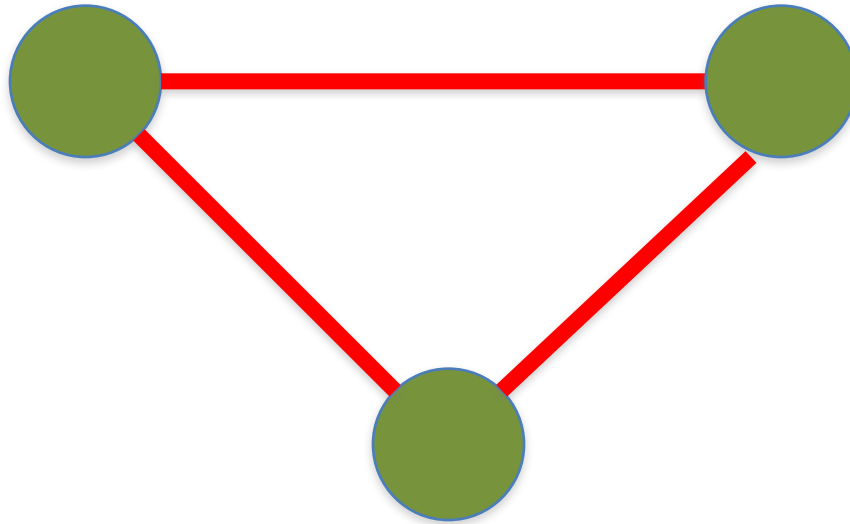


# Edge





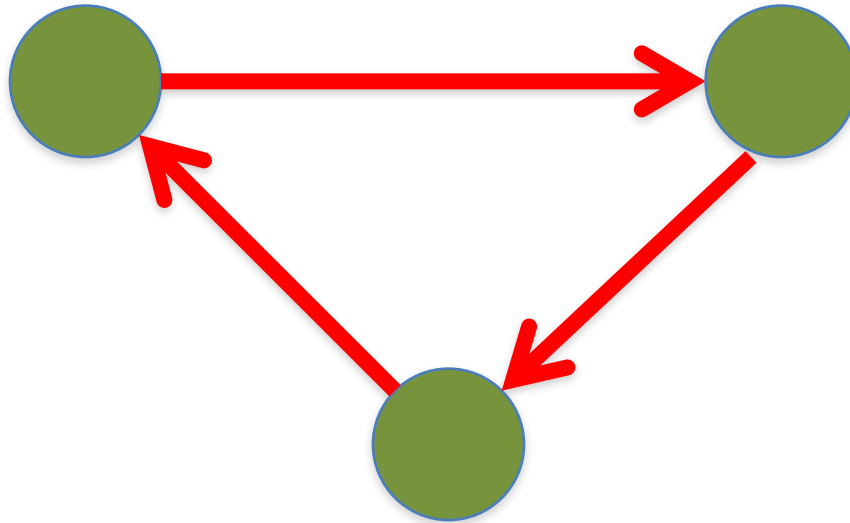
# Edges



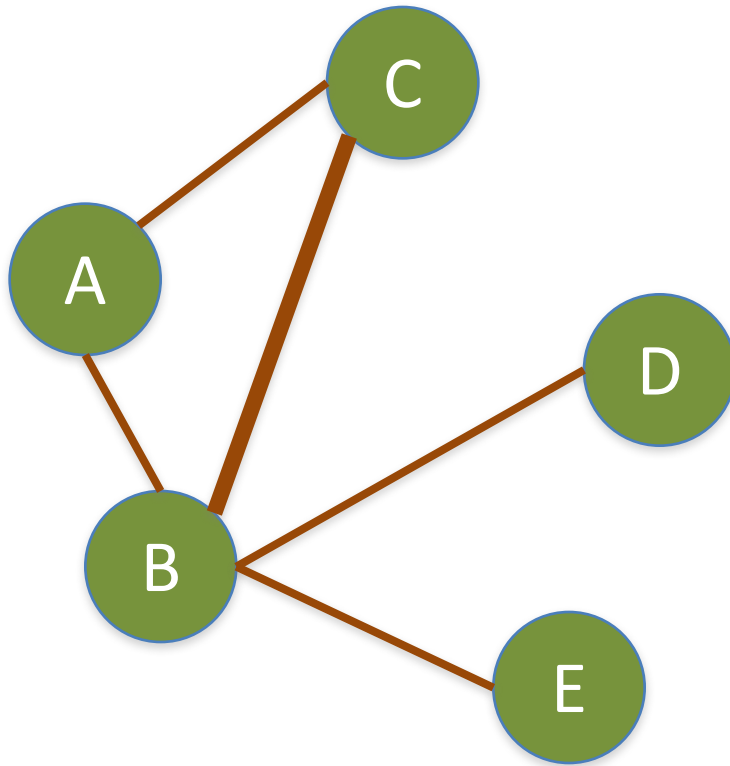
# Arc



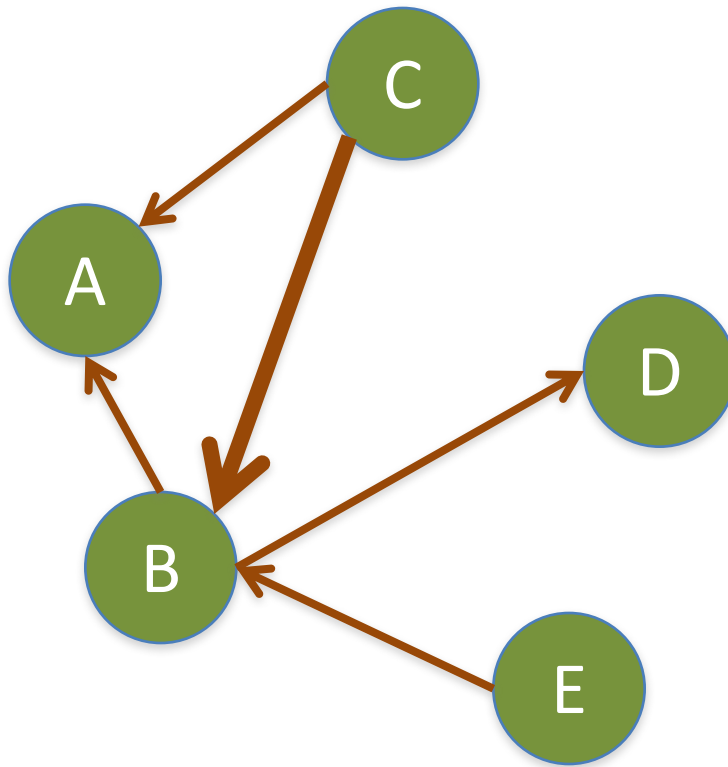
# Arcs



# Undirected Graph

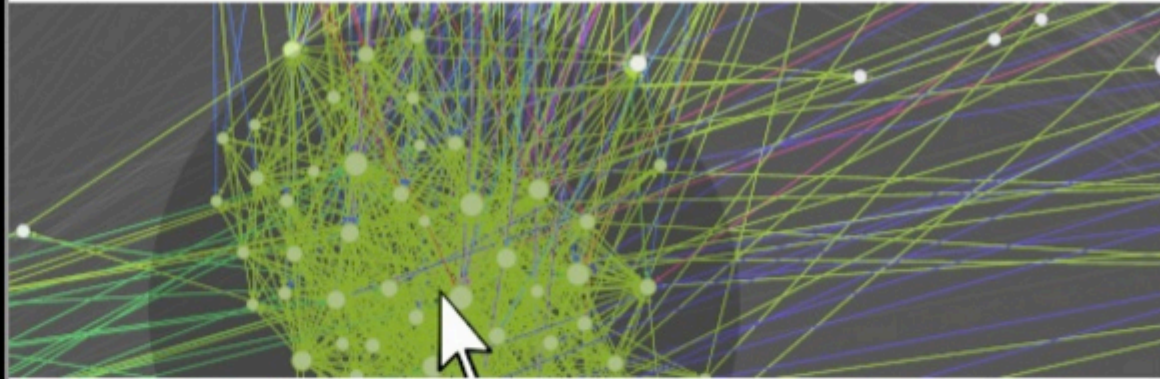
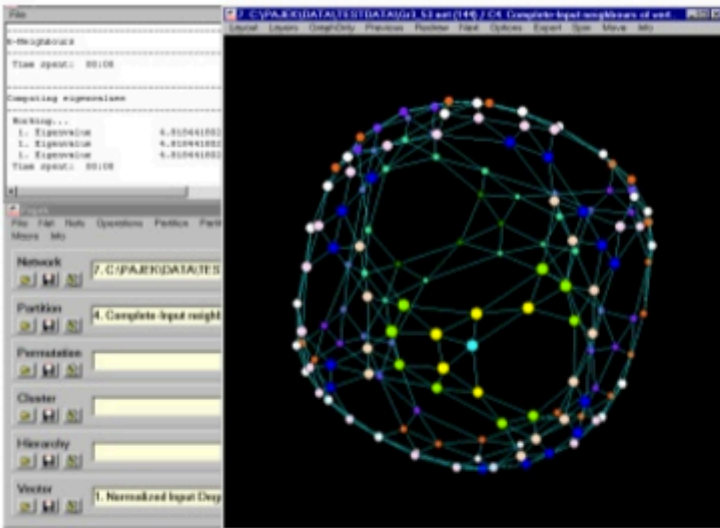


# Directed Graph



# Measurements of Social Network Analysis

# Exploratory Network Analysis



## 1 see the network

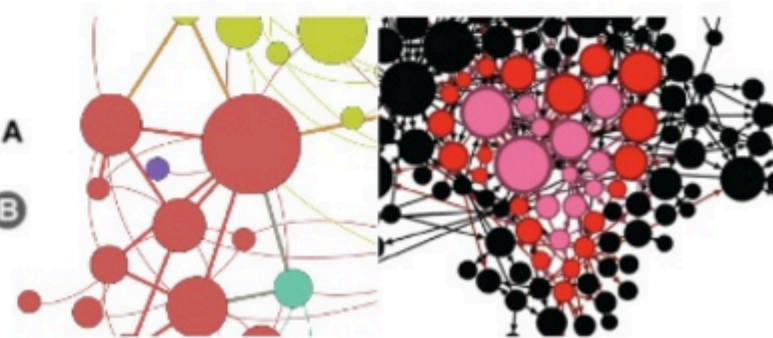
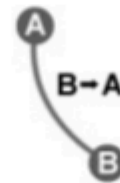
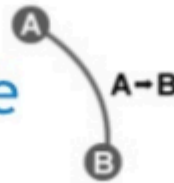
1st graph viz tool: Pajek (1996)  
Vladimir Batagelj, Andrej Mrvar

## 2 interact in real time

Gephi prototype (2008)  
group, filter, compute metrics...

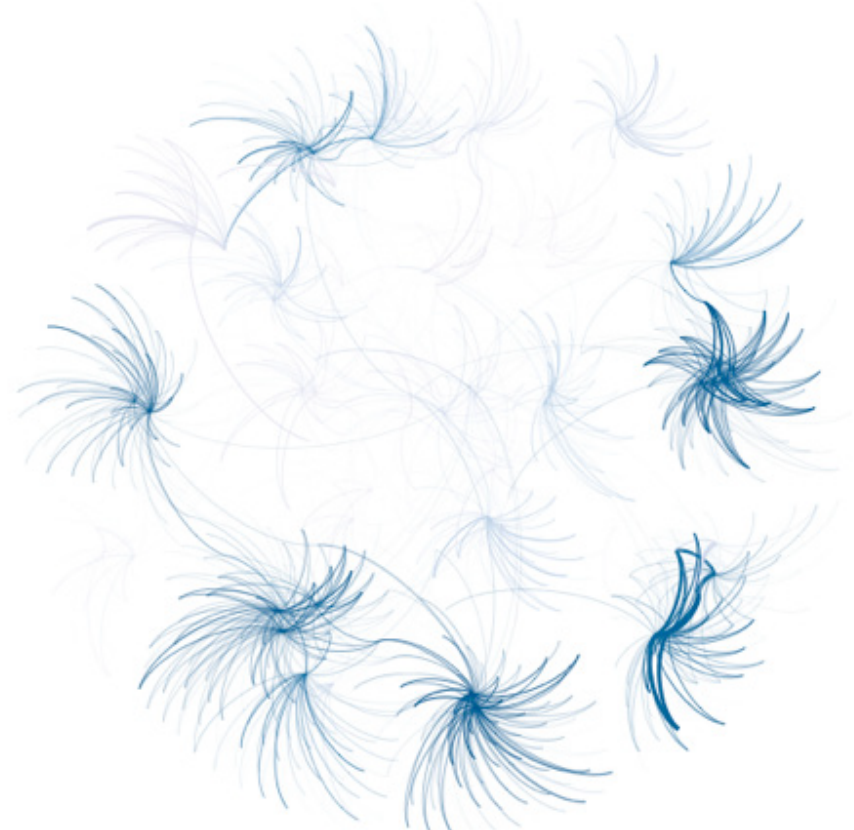
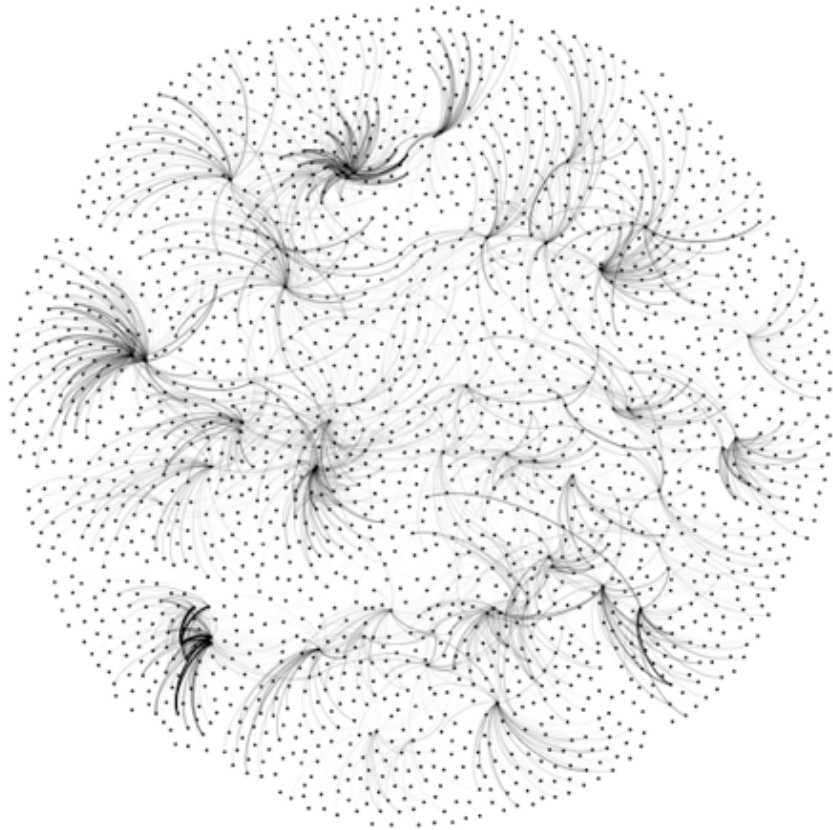
## 3 build a visual language

size by rank, color by partition,  
label, curved edges, thickness...



# Looking for a “Simple Small Truth”?

## What Data Visualization Should Do?



1. Make complex things **simple**
2. Extract **small** information from large data
3. Present **truth**, do not deceive



# Measurements

# Looking for Orderness in Data

Make varying 3 cursors simultaneously to extract **meaningful patterns**

MICRO level      MACRO level



*at different levels*

1 dimension      N dimensions



*on multiple dimensions*

T+0      T+N

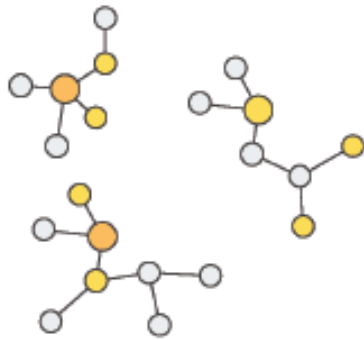


*at time scale*

# “Zoom” cursor on Quantitative Data

MICRO level

MACRO level



## Global

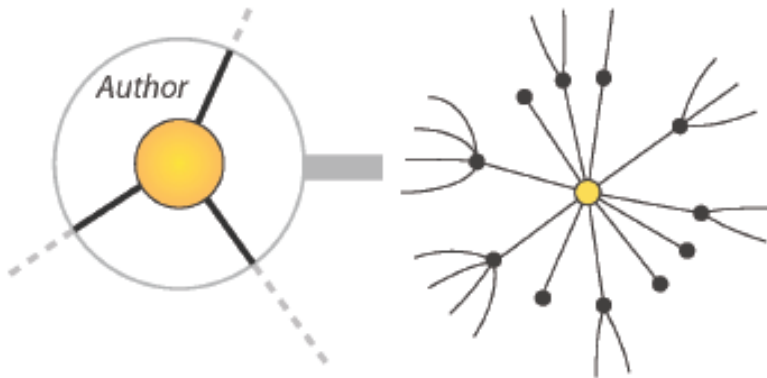
- connectivity
- density
- centralization

## Local

- communities
- bridges between communities
- local centers vs periphery

## Individual

- centrality
- distances
- neighborhood
- location
- local authority vs hub



# “Crossing” cursor on Quantitative Data



## Social

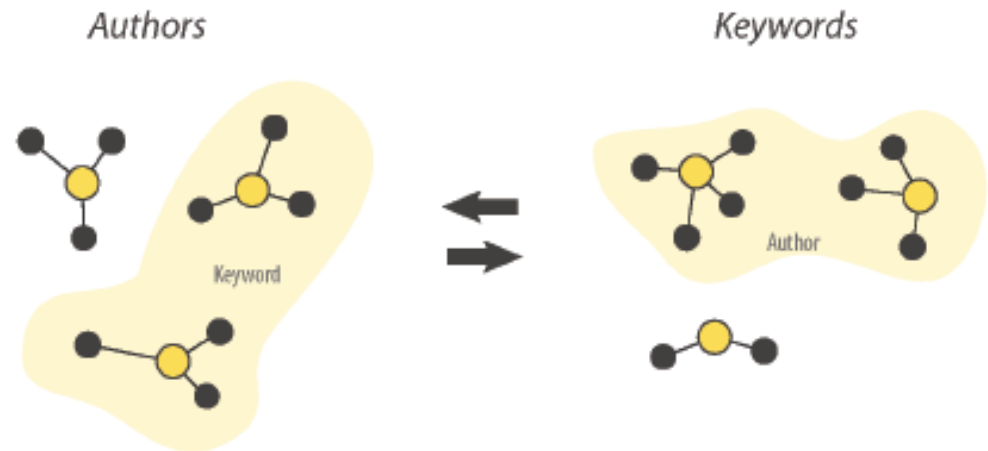
- who with whom
- communities
- brokerage
- influence and power
- homophily

## Semantic

- topics
- thematic clusters

## Geographic

- spatial phenomena



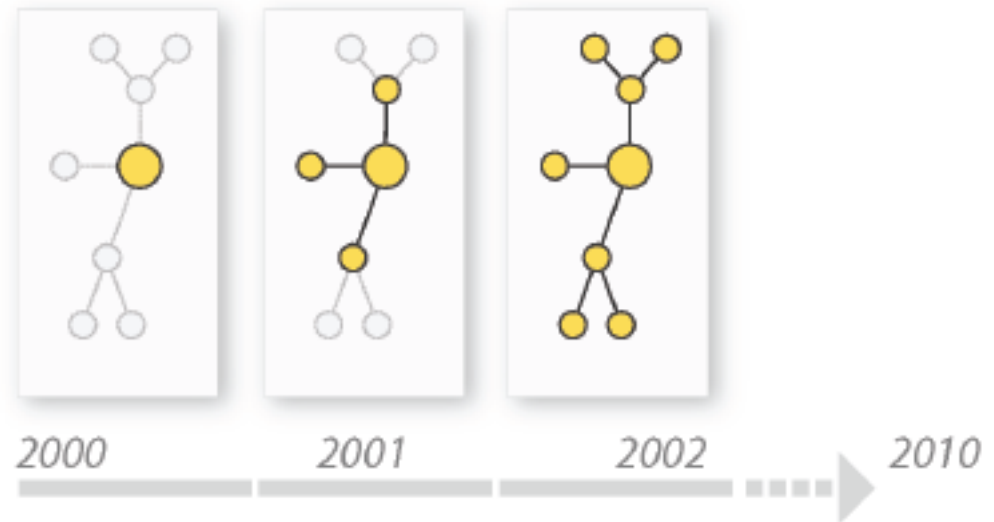
# “Timeline” cursor on Temporal Data



Evolution of social ties

Evolution of communities

Evolution of topics



# SNA Guideline

## # nodes

---

1 - 100

lists + edges in bonus, focus on qualitative data

100 - 1,000

### How attributes explain the structure?

- easy to read, “obvious” patterns
- focus on entities (in context)
- metrics are tools to describe the graph (centrality, bridging...)
- links help to build and interpret categories of entities

**challenge: mix attribute crossing and connectivity**

1,000 - 50,000

### How the structure explains attributes?

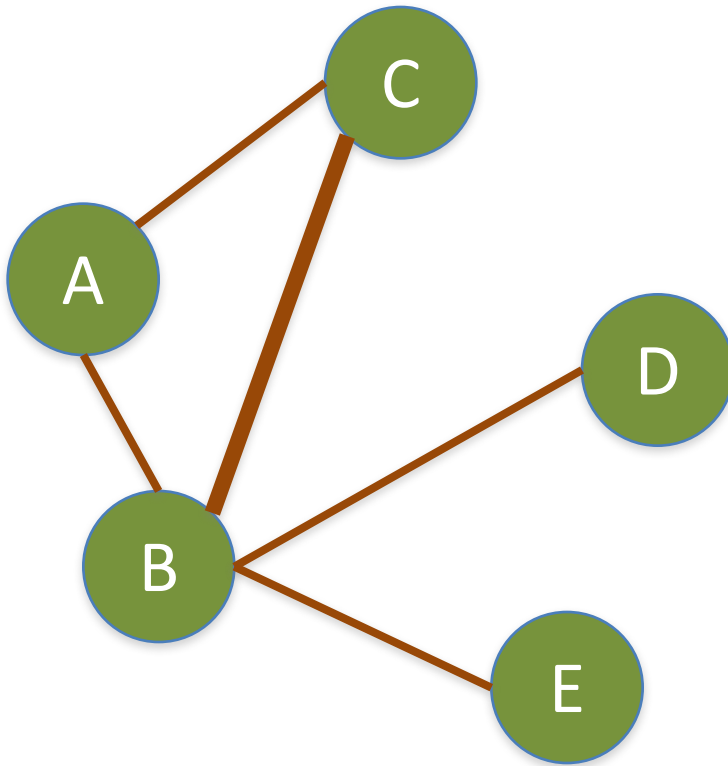
- hard to read, problem of “hidden signals”:  
track patterns with various layouts and filtering
- focus on structures
- metrics are tools to build the graph (cosine similarity...)
- categories help to understand the structure

**challenge: pattern recognition**

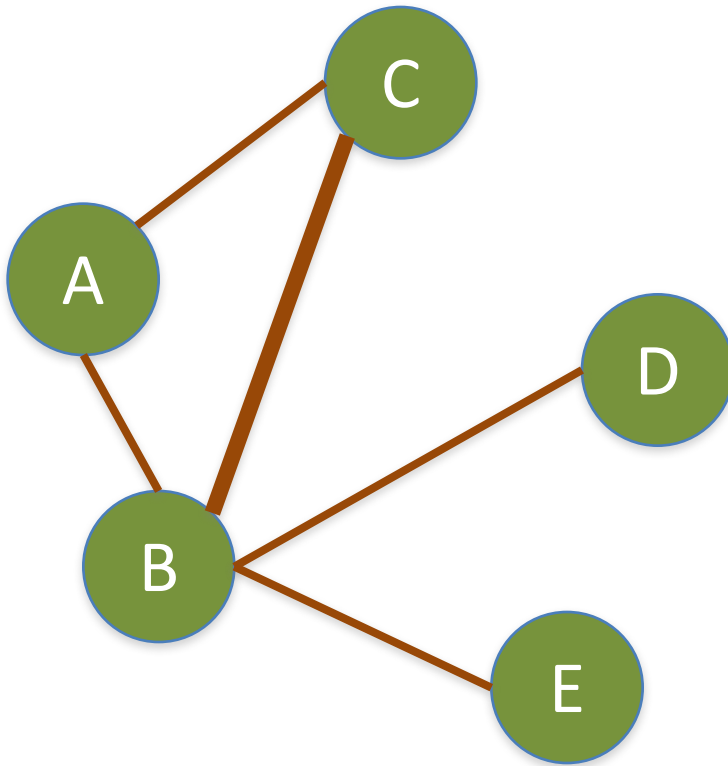
> 50,000

require high computational power

# Degree



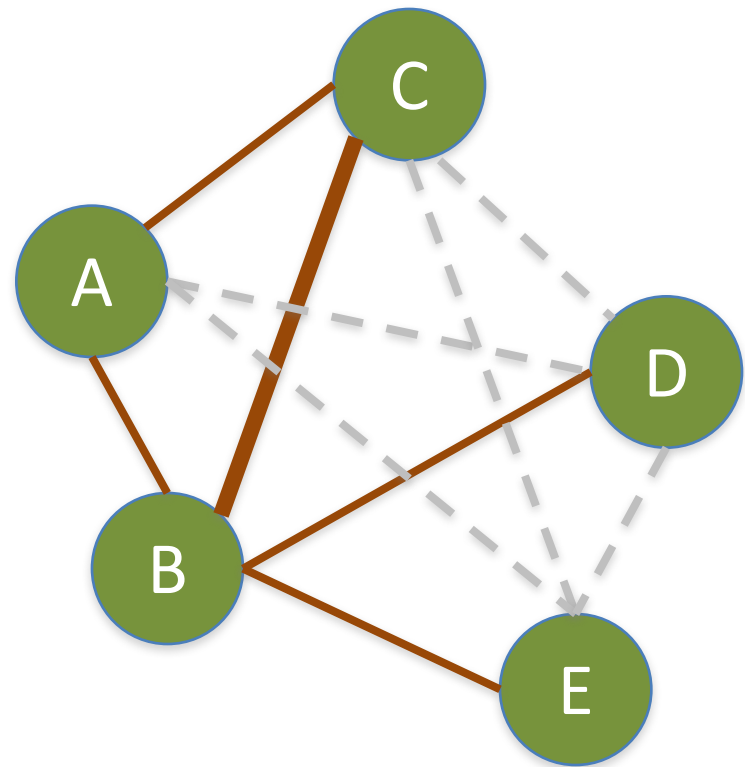
# Degree



A: 2  
B: 4  
C: 2  
D: 1  
E: 1



# Density

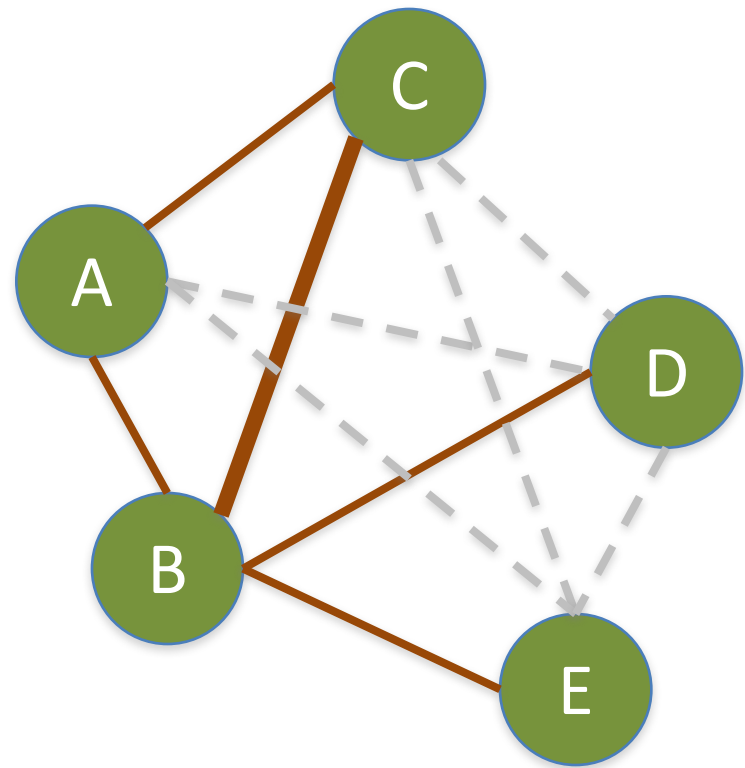


# Density

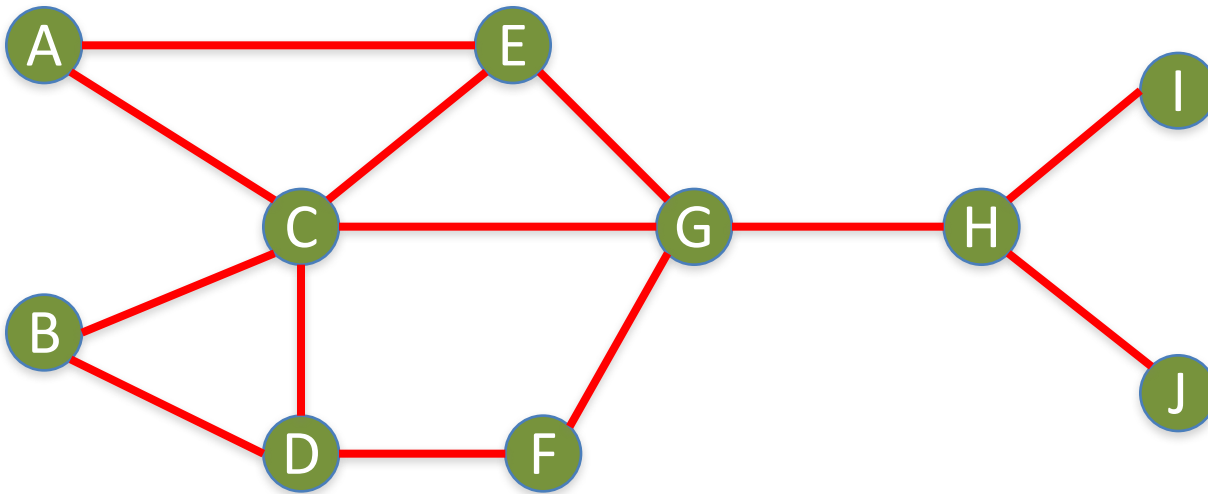
Edges (Links): 5

Total Possible Edges: 10

Density:  $5/10 = 0.5$



# Density



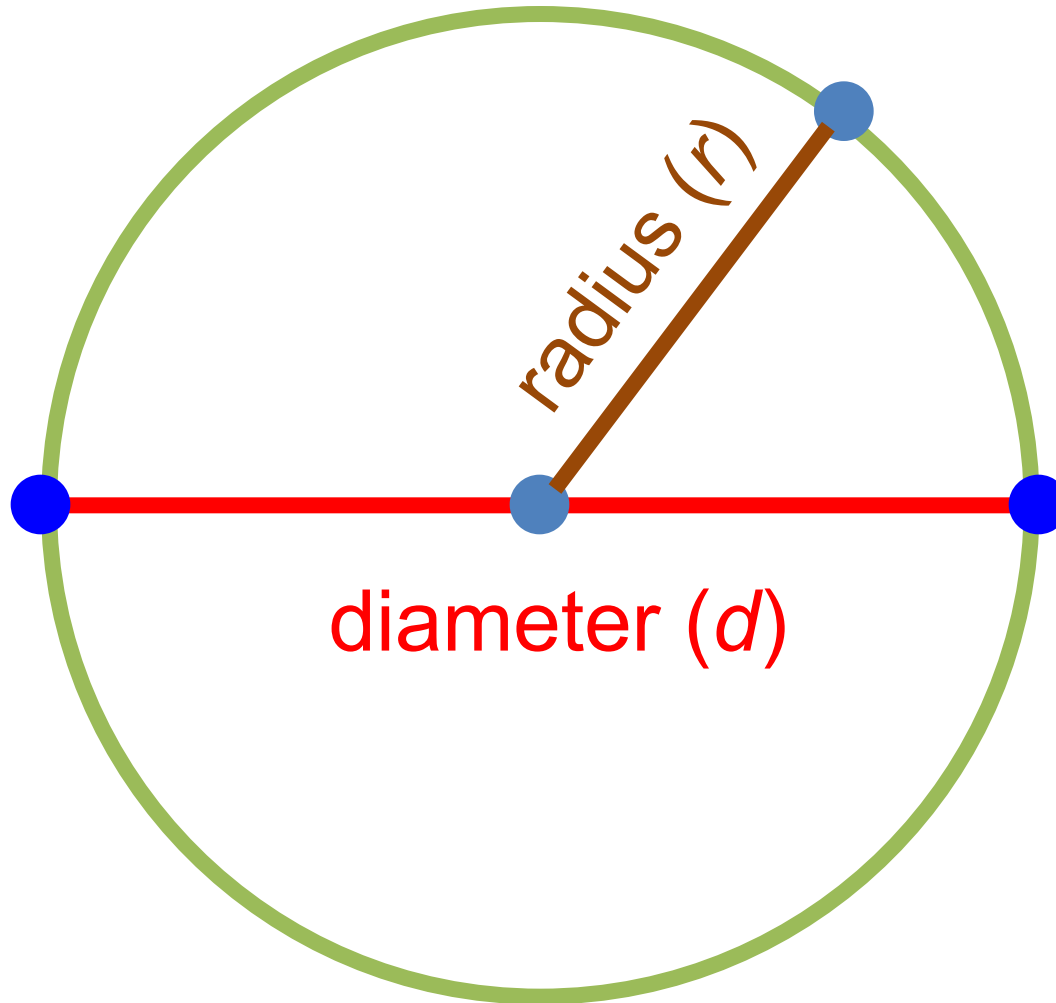
Nodes (n): 10

Edges (Links): 13

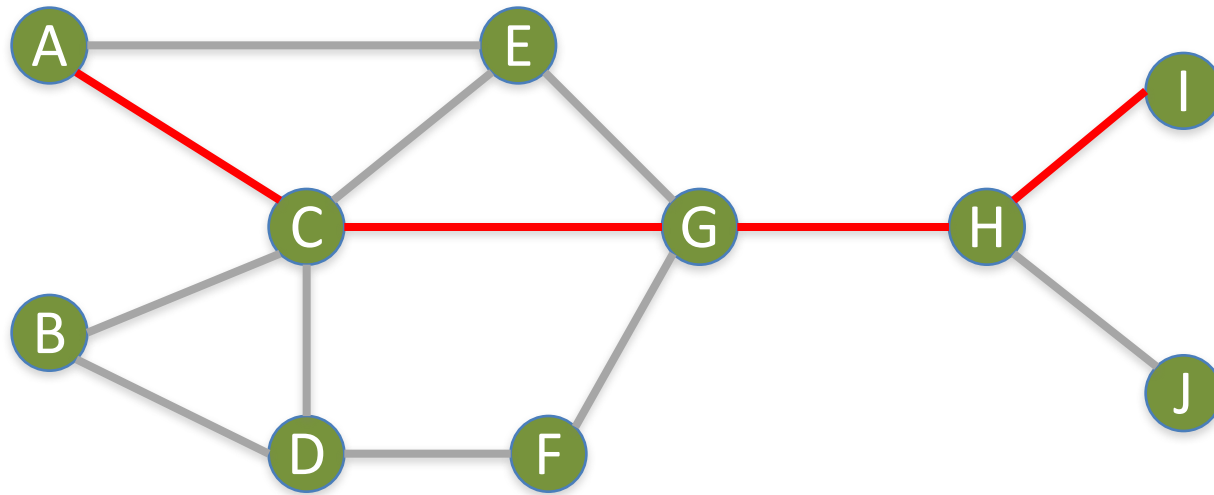
Total Possible Edges:  $(n * (n-1)) / 2 = (10 * 9) / 2 = 45$

Density:  $13/45 = 0.29$

# Diameter

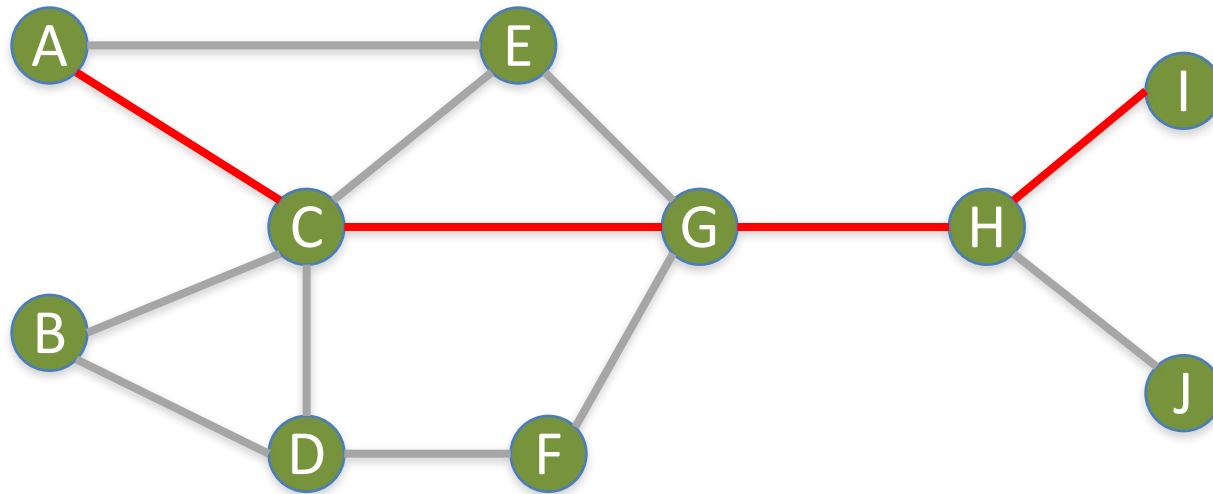


# Diameter



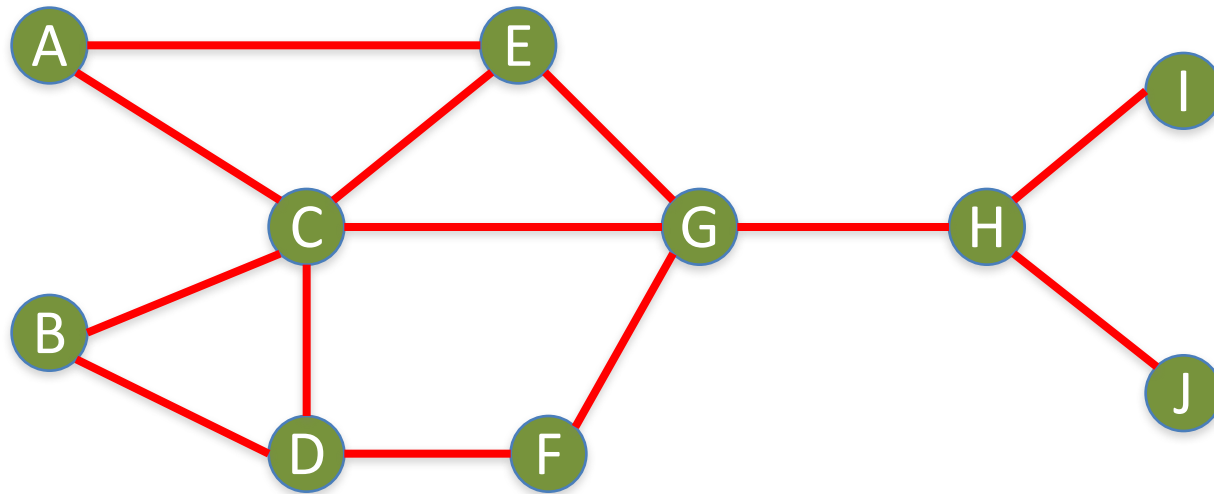
# Diameter

## Geodesic Path (Shortest Path)



**A → I : Diameter = 4**

# 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.

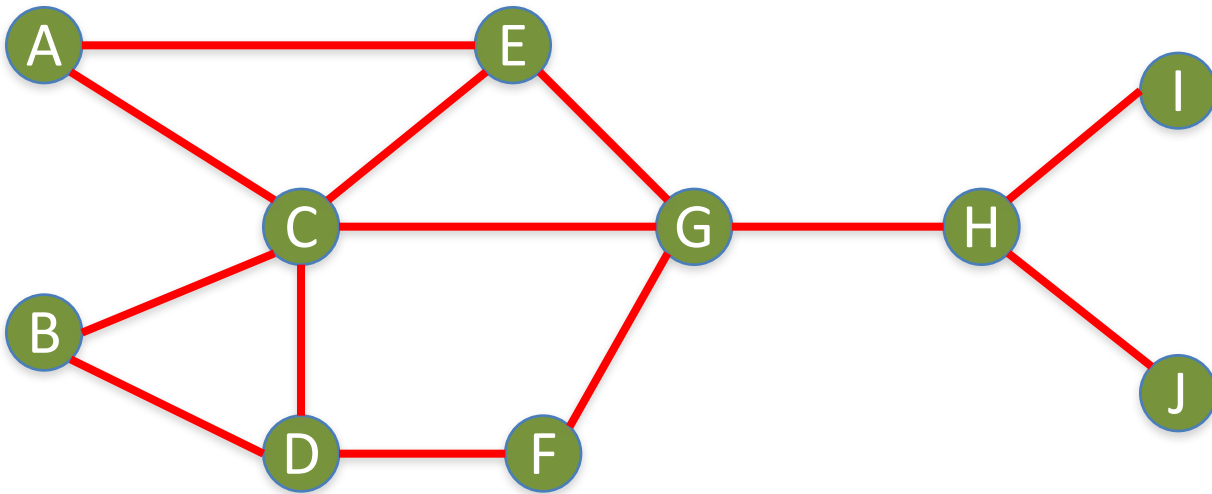


# Social Network Analysis (SNA)

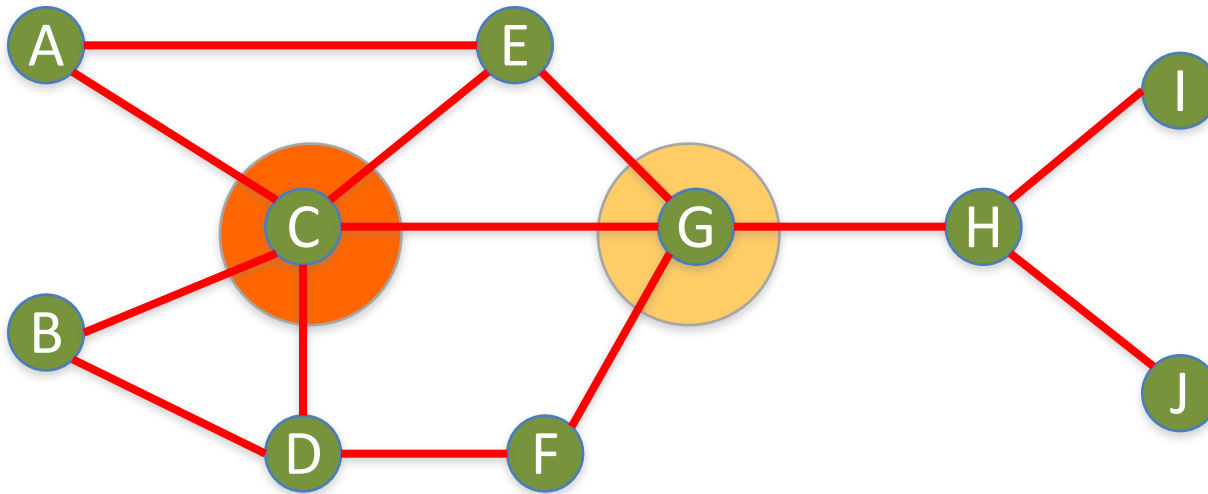
- Degree Centrality
- Betweenness Centrality
- Closeness Centrality

# Degree Centrality

# Social Network Analysis: Degree Centrality



# Social Network Analysis: Degree Centrality



Node	Score	Standardized Score
A	2	$2/10 = 0.2$
B	2	$2/10 = 0.2$
<b>C</b>	<b>5</b>	<b><math>5/10 = 0.5</math></b>
D	3	$3/10 = 0.3$
E	3	$3/10 = 0.3$
F	2	$2/10 = 0.2$
<b>G</b>	<b>4</b>	<b><math>4/10 = 0.4</math></b>
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

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

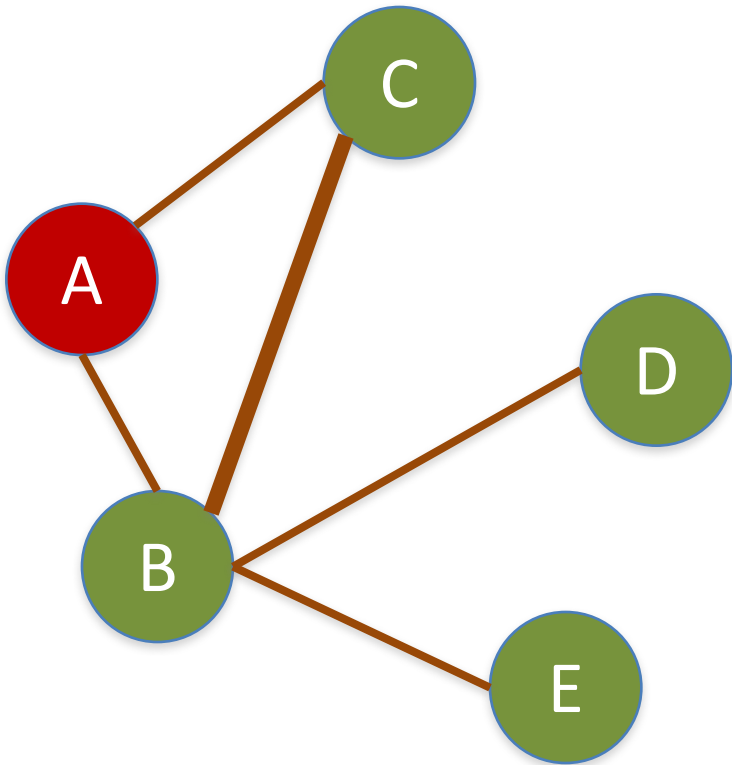
Where  $g_{jk}$  = the number of shortest paths connecting  $jk$   
 $g_{jk}(i)$  = the number that actor  $i$  is on.

## Normalized Betweenness Centrality

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

**Number of pairs of vertices  
excluding the vertex itself**

# Betweenness Centrality



A:

$$B \rightarrow C: 0/1 = 0$$

$$B \rightarrow D: 0/1 = 0$$

$$B \rightarrow E: 0/1 = 0$$

$$C \rightarrow D: 0/1 = 0$$

$$C \rightarrow E: 0/1 = 0$$

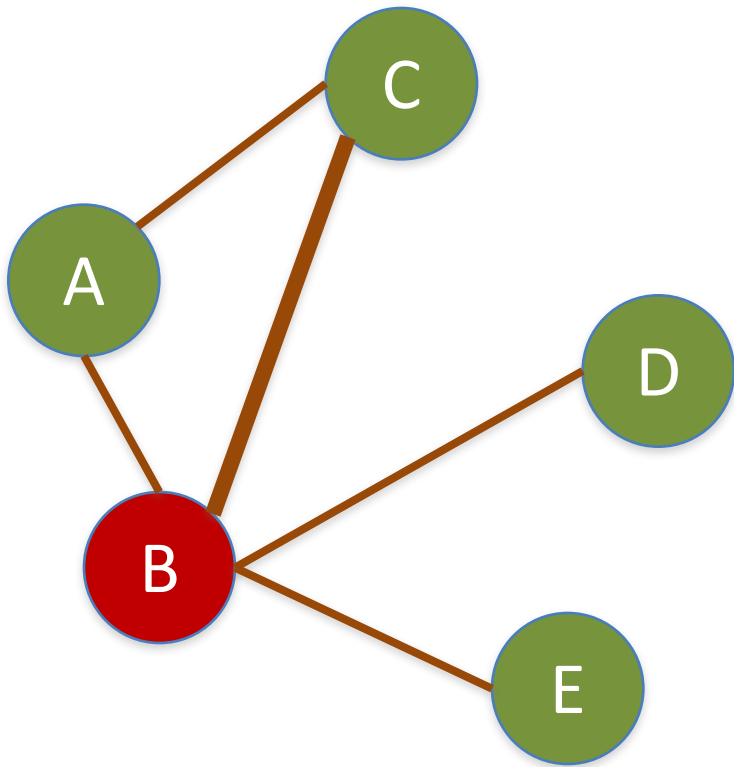
$$D \rightarrow E: 0/1 = 0$$

**Total:** 0

**A: Betweenness Centrality = 0**



# Betweenness Centrality



B:

$$A \rightarrow C: 0/1 = 0$$

$$A \rightarrow D: 1/1 = 1$$

$$A \rightarrow E: 1/1 = 1$$

$$C \rightarrow D: 1/1 = 1$$

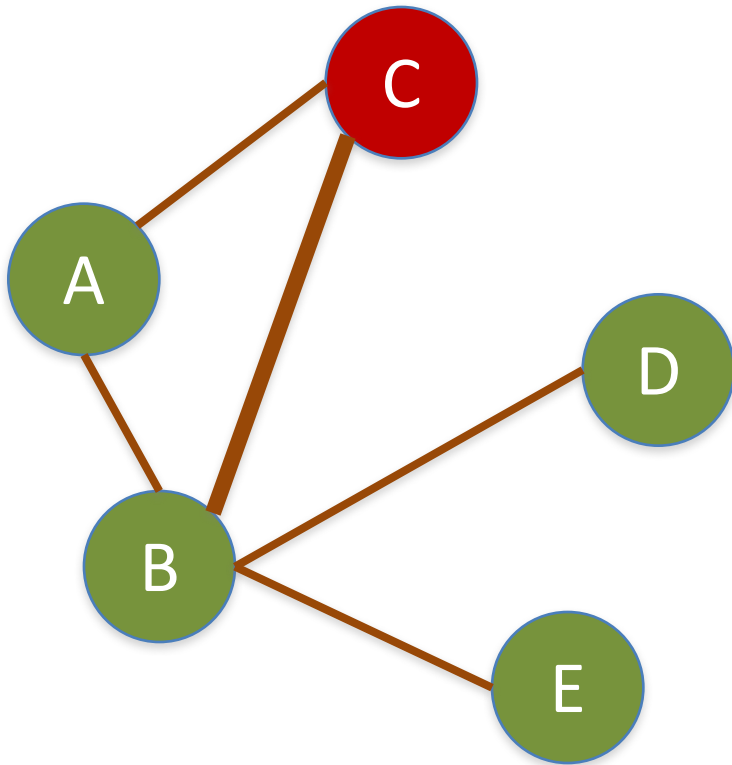
$$C \rightarrow E: 1/1 = 1$$

$$D \rightarrow E: 1/1 = 1$$

**Total:** 5

**B: Betweenness Centrality = 5**

# Betweenness Centrality



C:

$$A \rightarrow B: 0/1 = 0$$

$$A \rightarrow D: 0/1 = 0$$

$$A \rightarrow E: 0/1 = 0$$

$$B \rightarrow D: 0/1 = 0$$

$$B \rightarrow E: 0/1 = 0$$

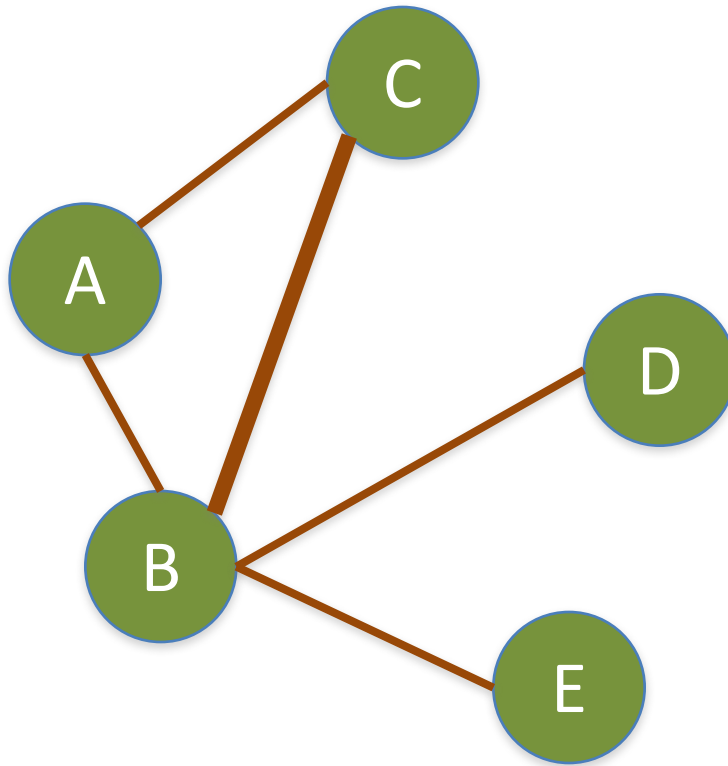
$$D \rightarrow E: 0/1 = 0$$

---

$$\text{Total: } \quad \quad \quad \underline{\quad 0 \quad}$$

**C: Betweenness Centrality = 0**

# Betweenness Centrality



A: 0

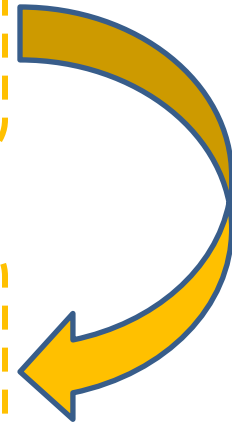
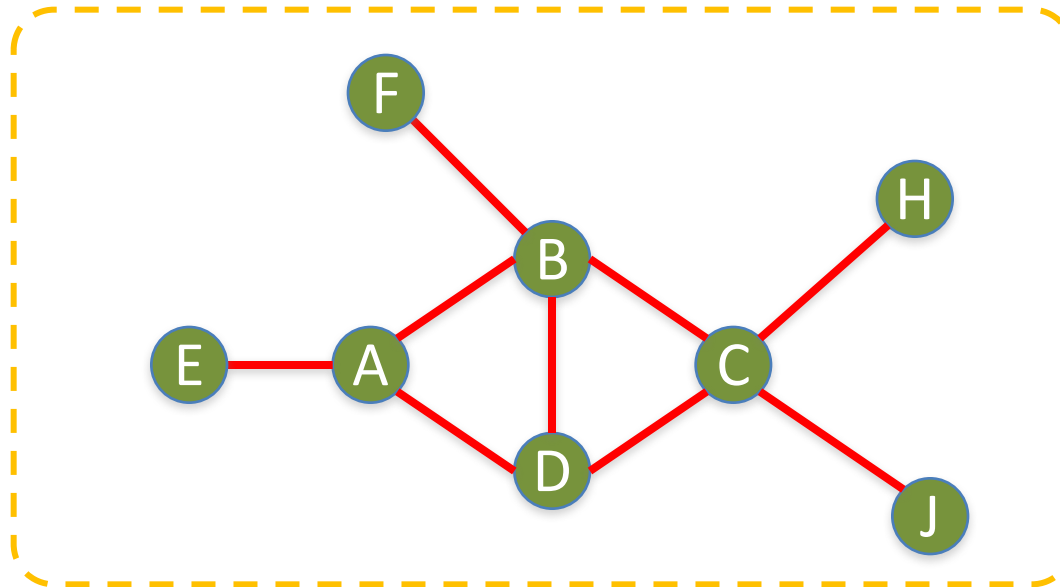
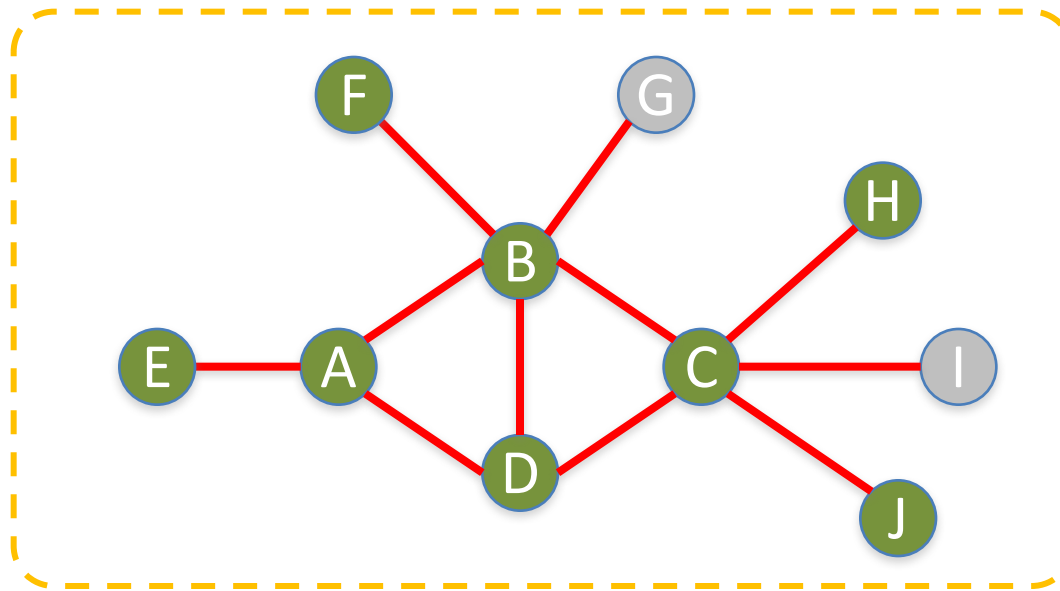
**B: 5**

C: 0

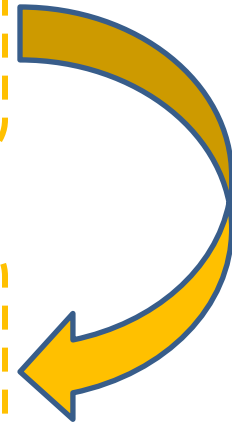
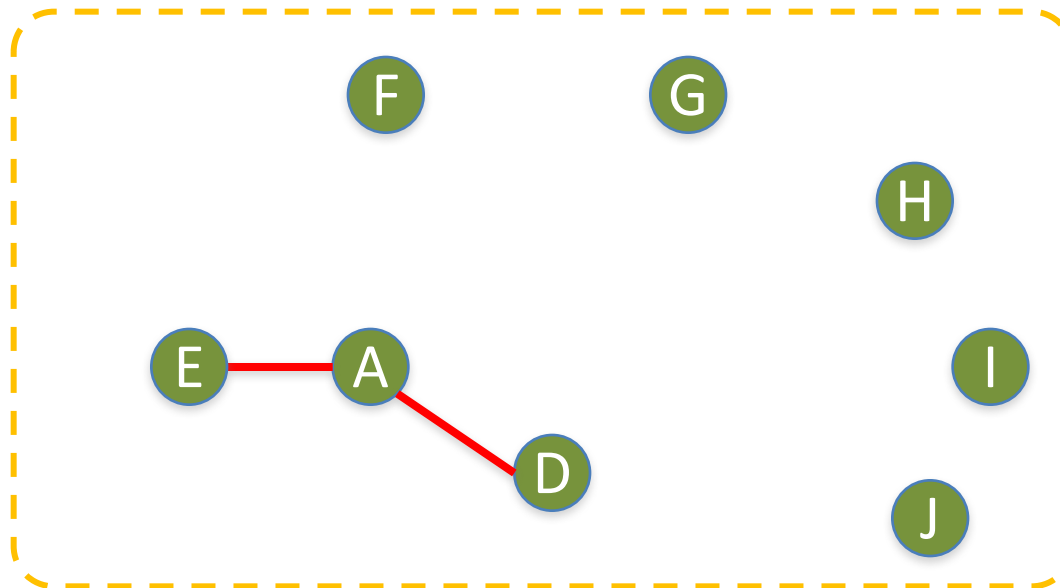
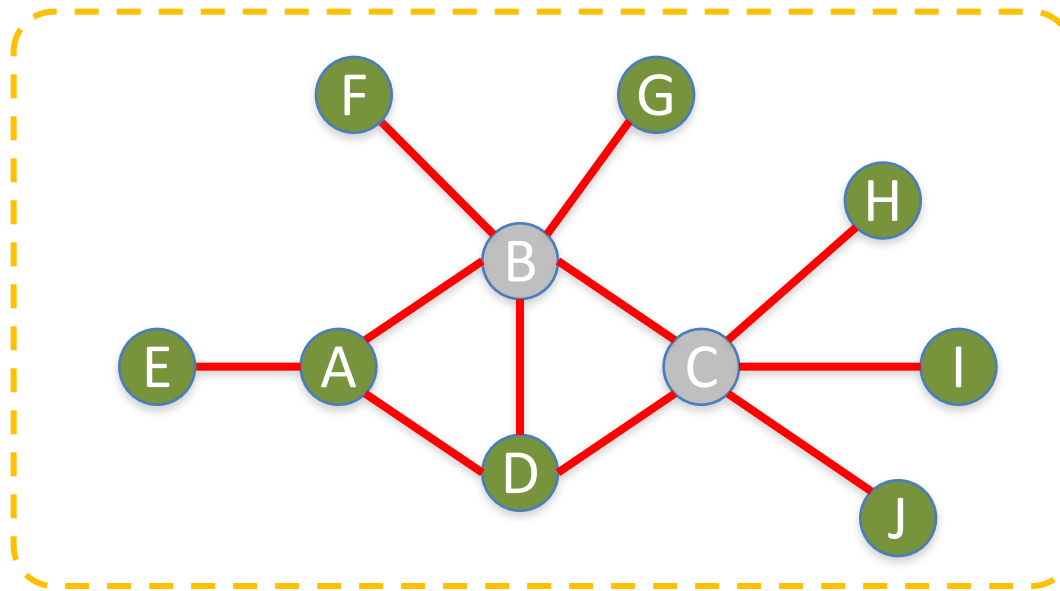
D: 0

E: 0

# Which Node is Most **Important**?

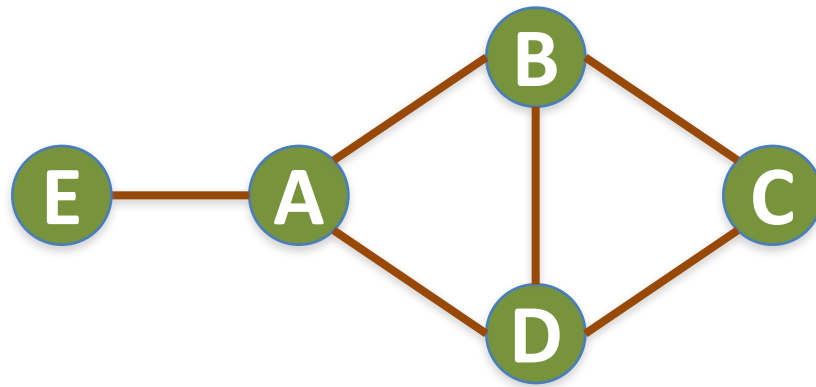


# Which Node is Most **Important**?

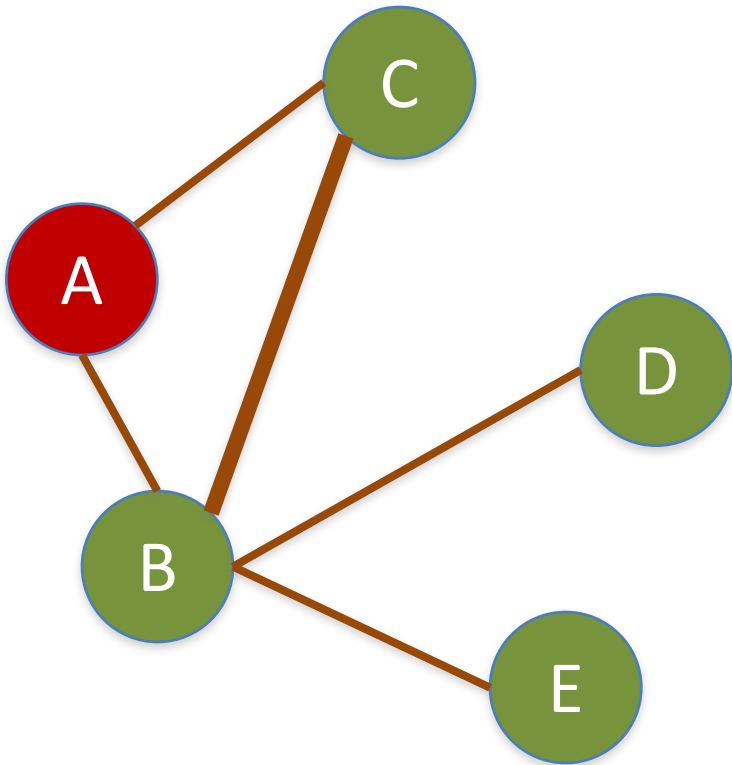


# Betweenness Centrality

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



# Betweenness Centrality



A:

$$B \rightarrow C: 0/1 = 0$$

$$B \rightarrow D: 0/1 = 0$$

$$B \rightarrow E: 0/1 = 0$$

$$C \rightarrow D: 0/1 = 0$$

$$C \rightarrow E: 0/1 = 0$$

$$D \rightarrow E: 0/1 = 0$$

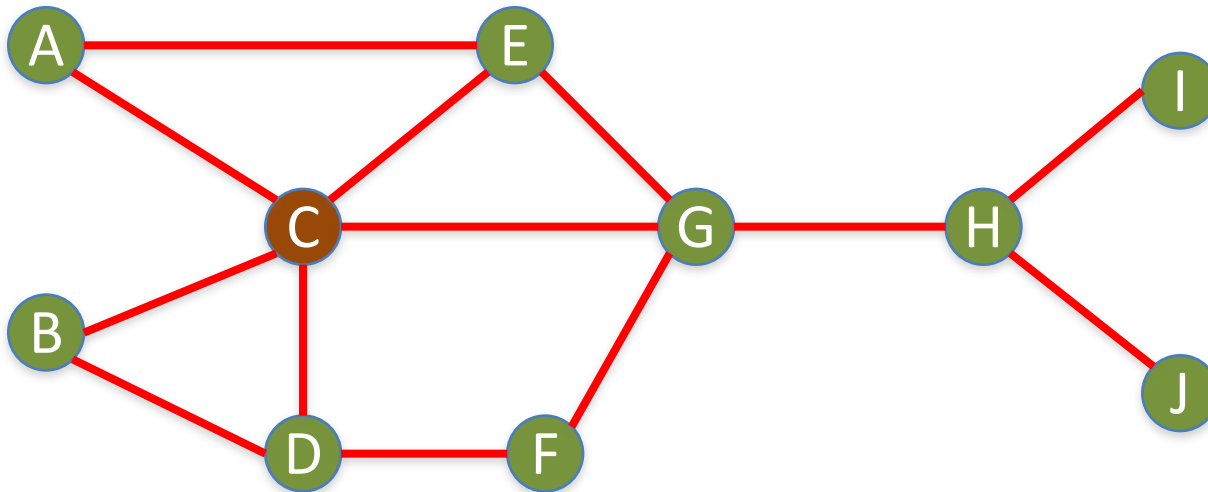
**Total:** 0

**A: Betweenness Centrality = 0**

**Closeness**  
**Centrality**



# Social Network Analysis: Closeness Centrality



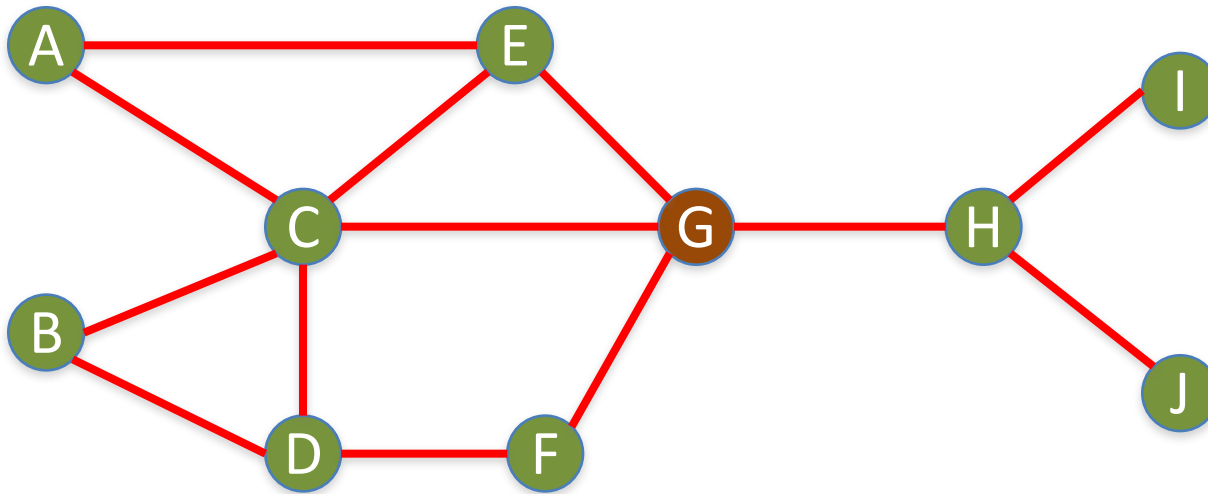
C→A: 1  
C→B: 1  
C→D: 1  
C→E: 1  
C→F: 2  
C→G: 1  
C→H: 2  
C→I: 3  
C→J: 3

---

Total=15

**C: Closeness Centrality =  $15/9 = 1.67$**

# Social Network Analysis: Closeness Centrality



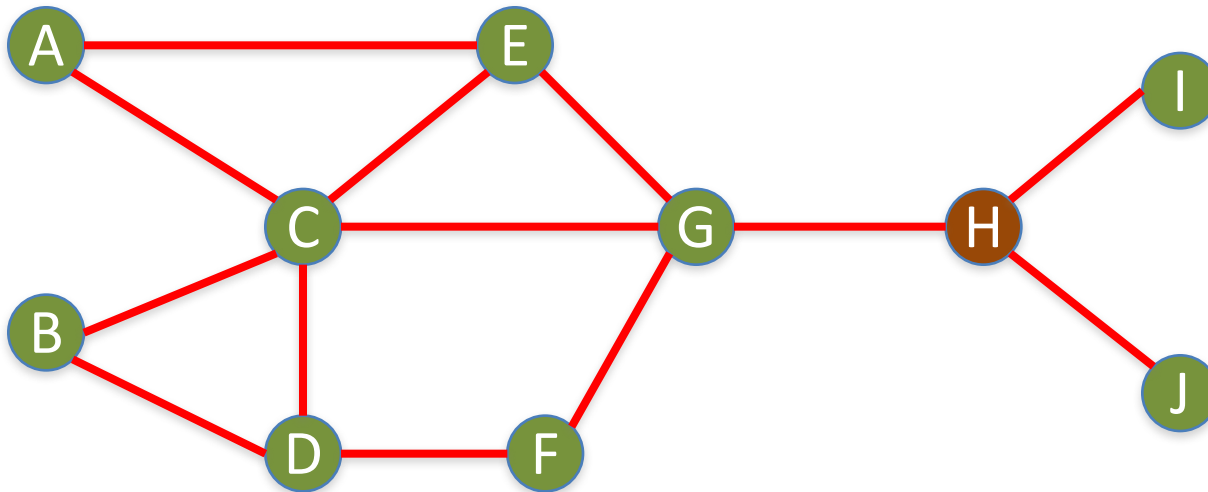
G→A: 2  
G→B: 2  
G→C: 1  
G→D: 2  
G→E: 1  
G→F: 1  
G→H: 1  
G→I: 2  
G→J: 2

---

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

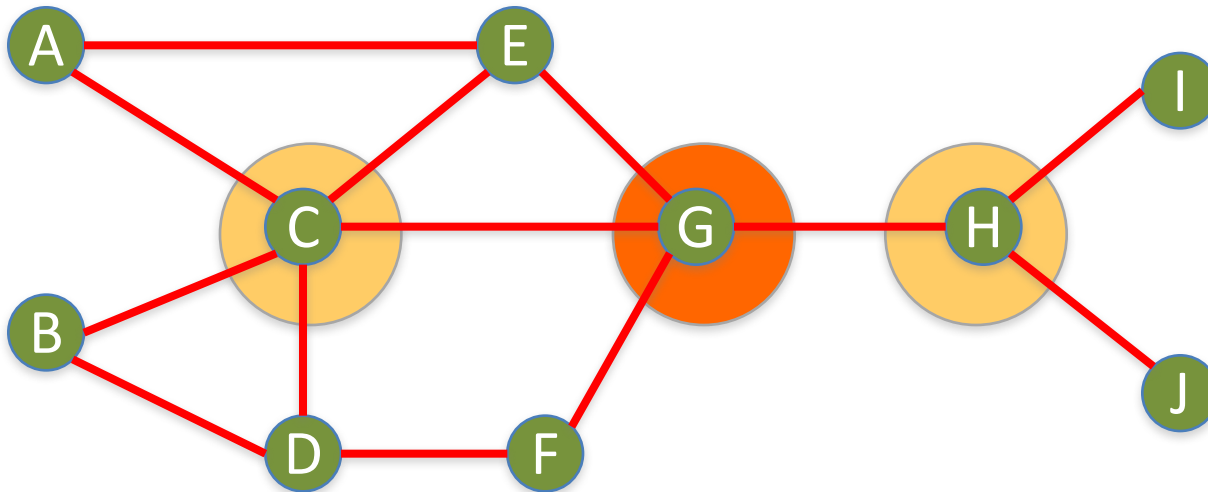
H→J: 1

---

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

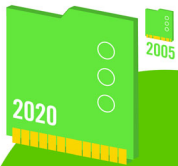
# Big Data Analytics

# Big Data 4 V

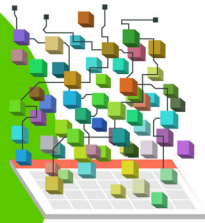
**40 ZETTABYTES**  
[ 43 TRILLION GIGABYTES ]  
of data will be created by 2020, an increase of 300 times from 2005



## Volume SCALE OF DATA



It's estimated that **2.5 QUINTILLION BYTES**  
[ 2.3 TRILLION GIGABYTES ]  
of data are created each day



Most companies in the U.S. have at least **100 TERABYTES**  
[ 100,000 GIGABYTES ]  
of data stored



## The FOUR V's of Big Data

From traffic patterns and music downloads to web history and medical records, data is recorded, stored, and analyzed to enable the technology and services that the world relies on every day. But what exactly is big data, and how can these massive amounts of data be used?

As a leader in the sector, IBM data scientists break big data into four dimensions: **Volume, Velocity, Variety and Veracity**

Depending on the industry and organization, big data encompasses information from multiple internal and external sources such as transactions, social media, enterprise content, sensors and mobile devices. Companies can leverage data to adapt their products and services to better meet customer needs, optimize operations and infrastructure, and find new sources of revenue.

By 2015 **4.4 MILLION IT JOBS** will be created globally to support big data, with 1.9 million in the United States



As of 2011, the global size of data in healthcare was estimated to be **150 EXABYTES**  
[ 161 BILLION GIGABYTES ]



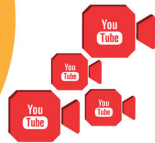
**30 BILLION PIECES OF CONTENT** are shared on Facebook every month



By 2014, it's anticipated there will be **420 MILLION WEARABLE, WIRELESS HEALTH MONITORS**

## Variety DIFFERENT FORMS OF DATA

**4 BILLION+ HOURS OF VIDEO** are watched on YouTube each month



**400 MILLION TWEETS** are sent per day by about 200 million monthly active users



The New York Stock Exchange captures **1 TB OF TRADE INFORMATION** during each trading session



## Velocity ANALYSIS OF STREAMING DATA



Modern cars have close to **100 SENSORS** that monitor items such as fuel level and tire pressure

By 2016, it is projected there will be **18.9 BILLION NETWORK CONNECTIONS** – almost 2.5 connections per person on earth



**1 IN 3 BUSINESS LEADERS** don't trust the information they use to make decisions



**27% OF RESPONDENTS**

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

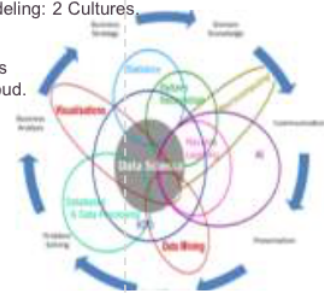
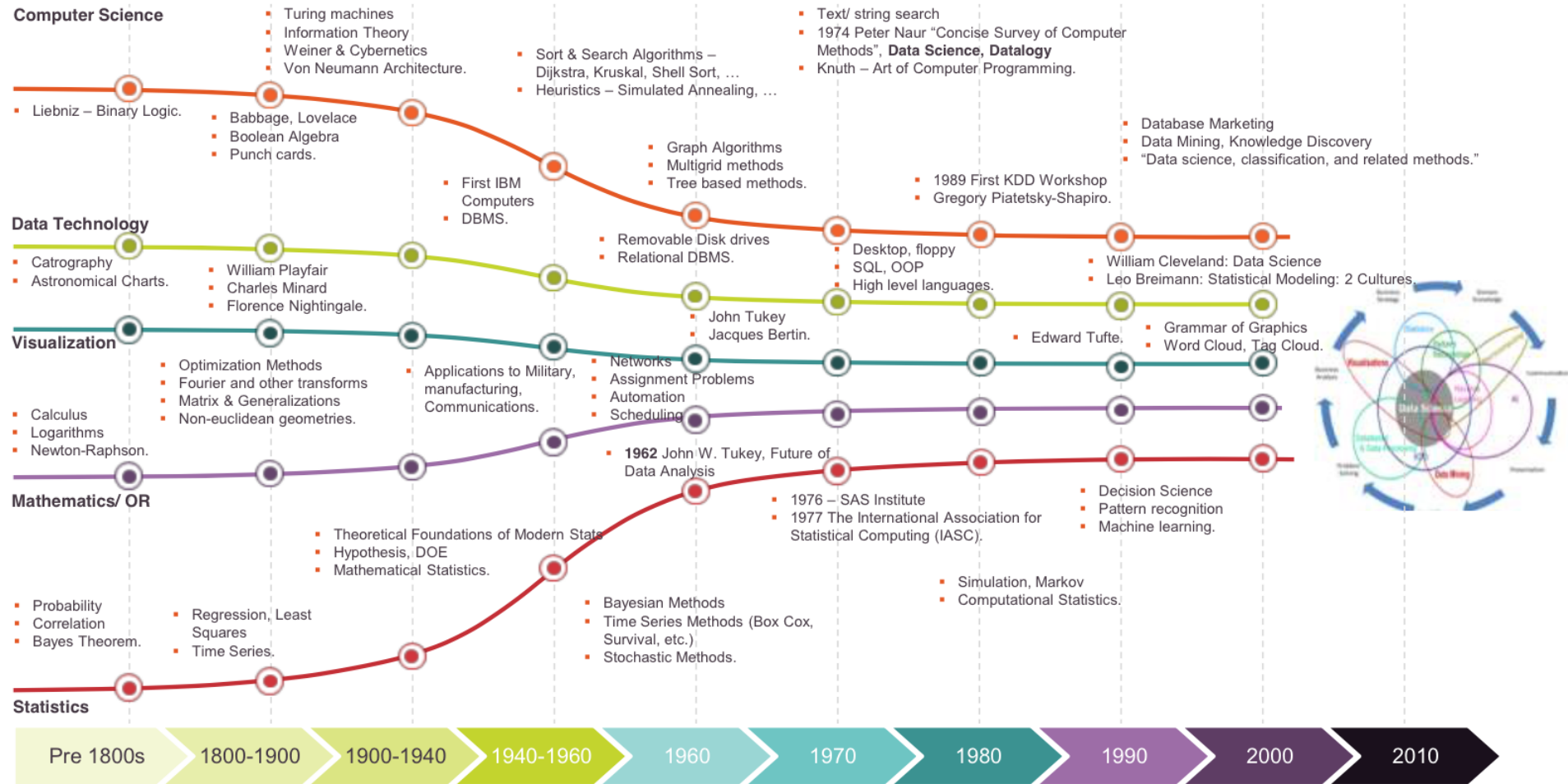
## Veracity UNCERTAINTY OF DATA

Poor data quality costs the US economy around **\$3.1 TRILLION A YEAR**



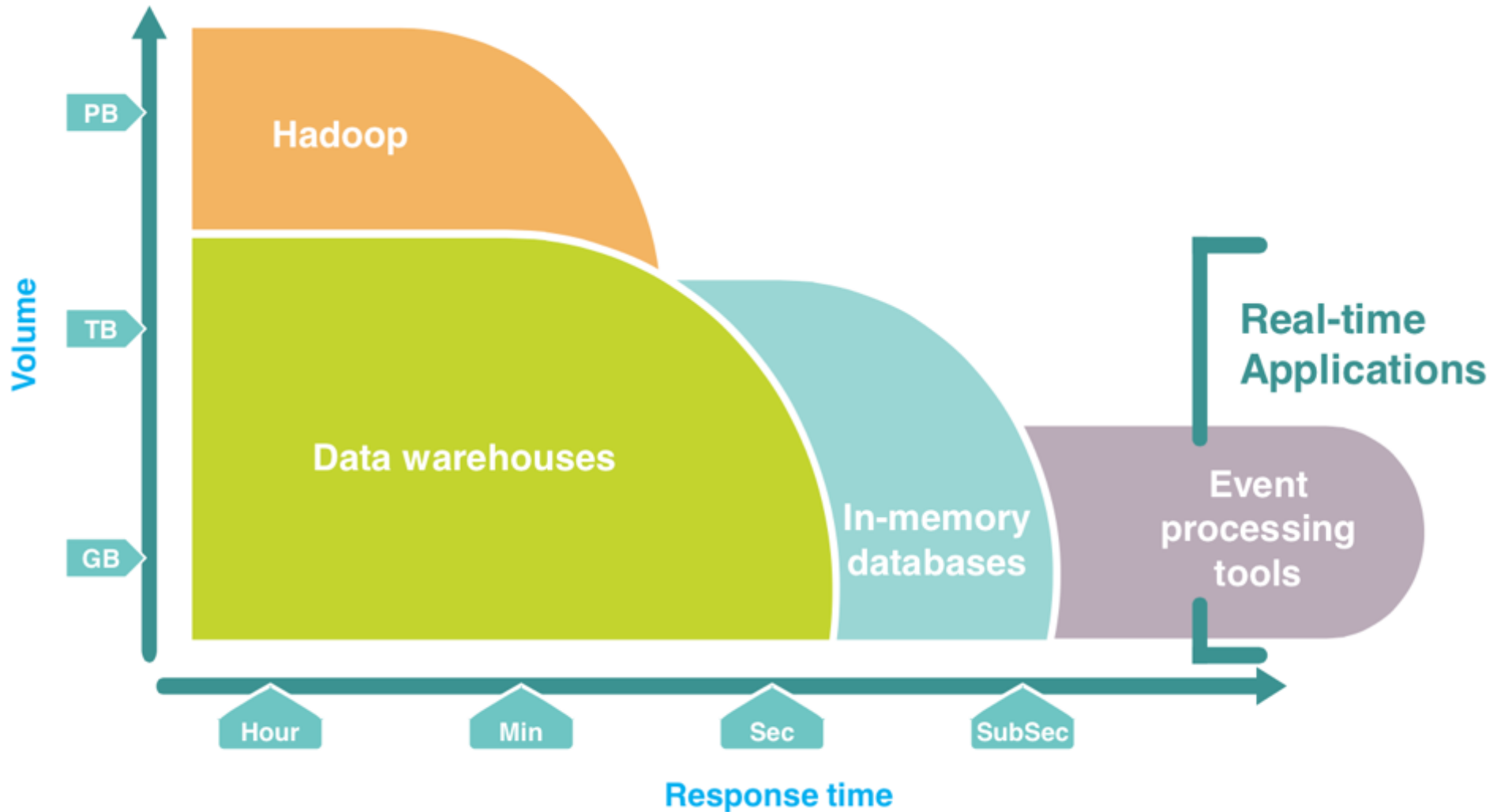
**value**

# History of Data Science



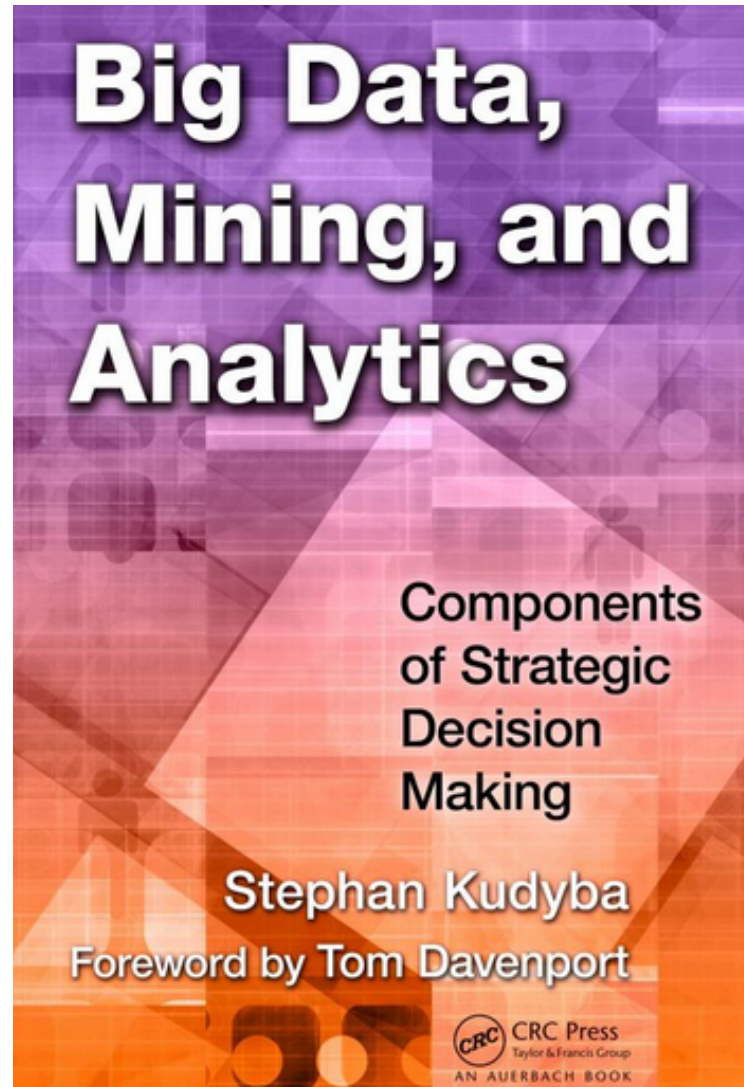


# Big Data Technologies are Enabling a New Approach

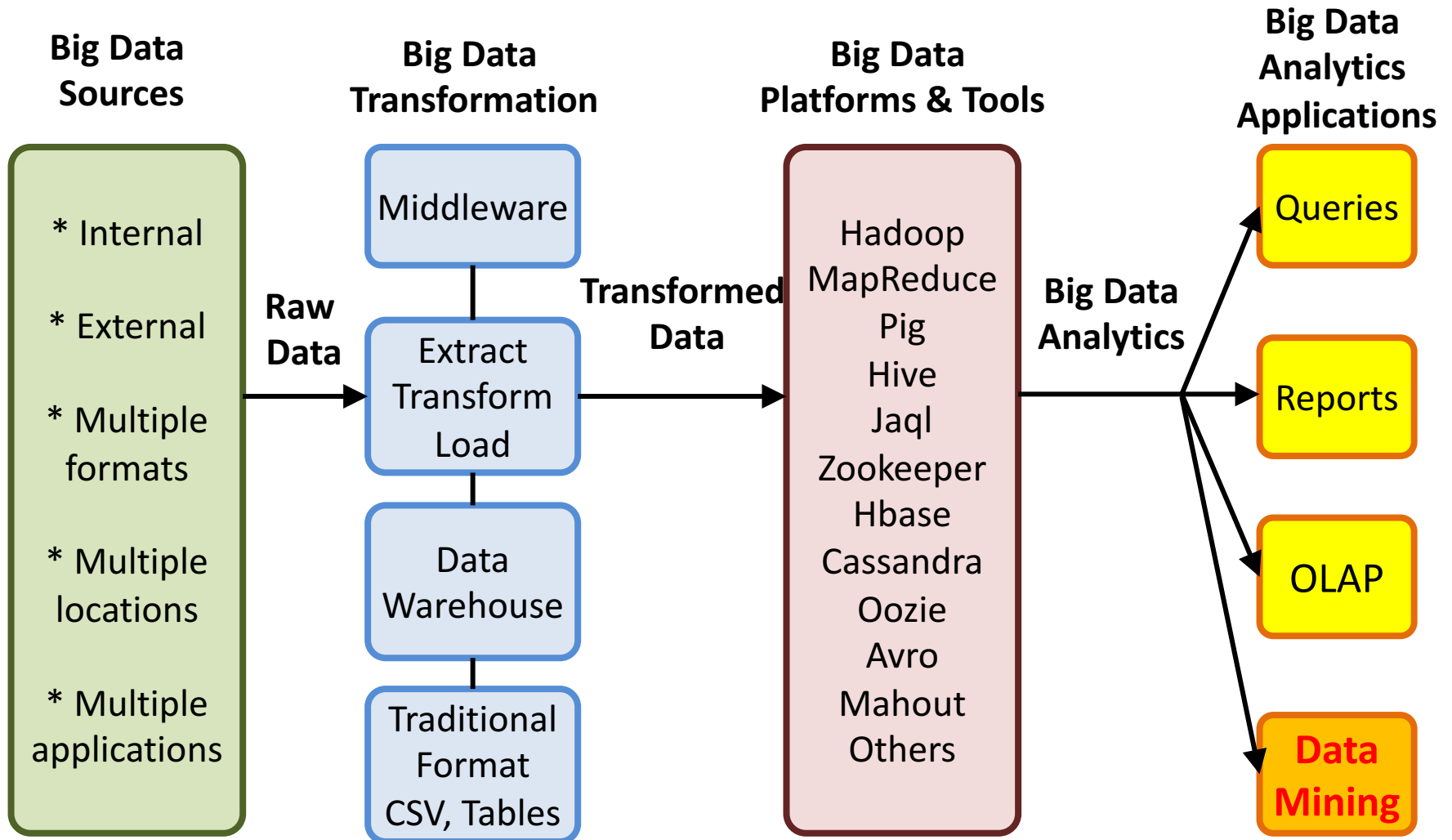


**Big Data**  
**Analytics**  
and  
**Data Mining**

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



# Architecture of Big Data Analytics

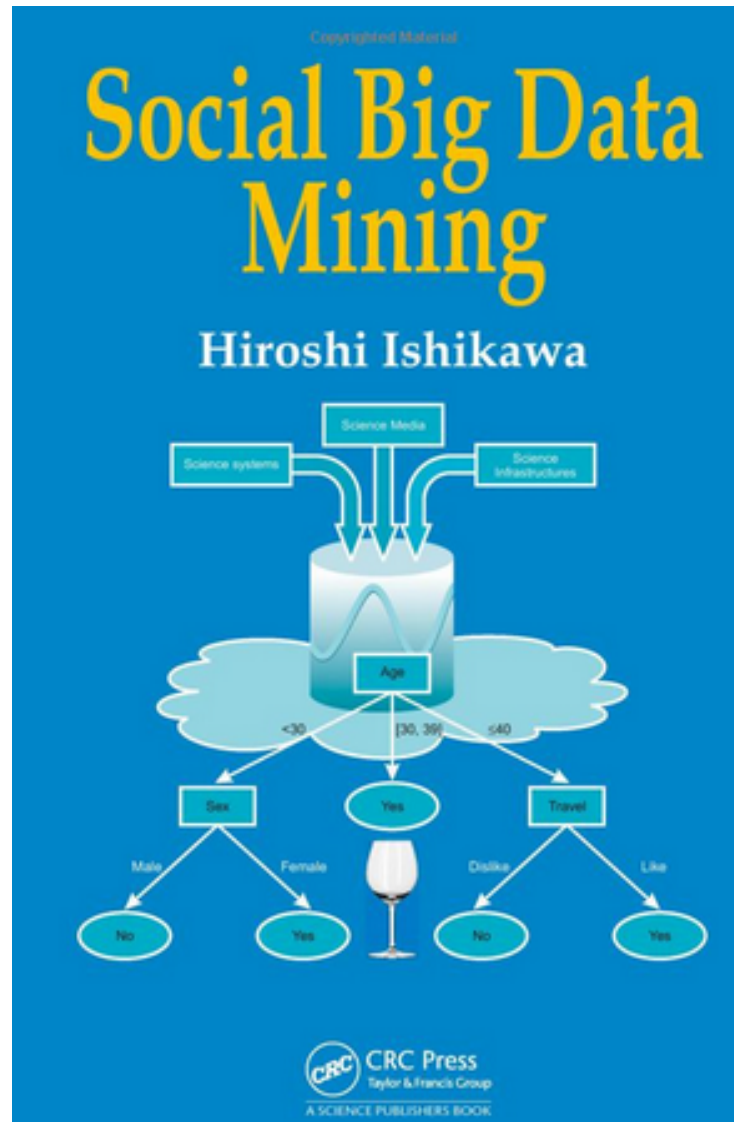


# Architecture of Big Data Analytics



# Social Big Data Mining

(Hiroshi Ishikawa, 2015)

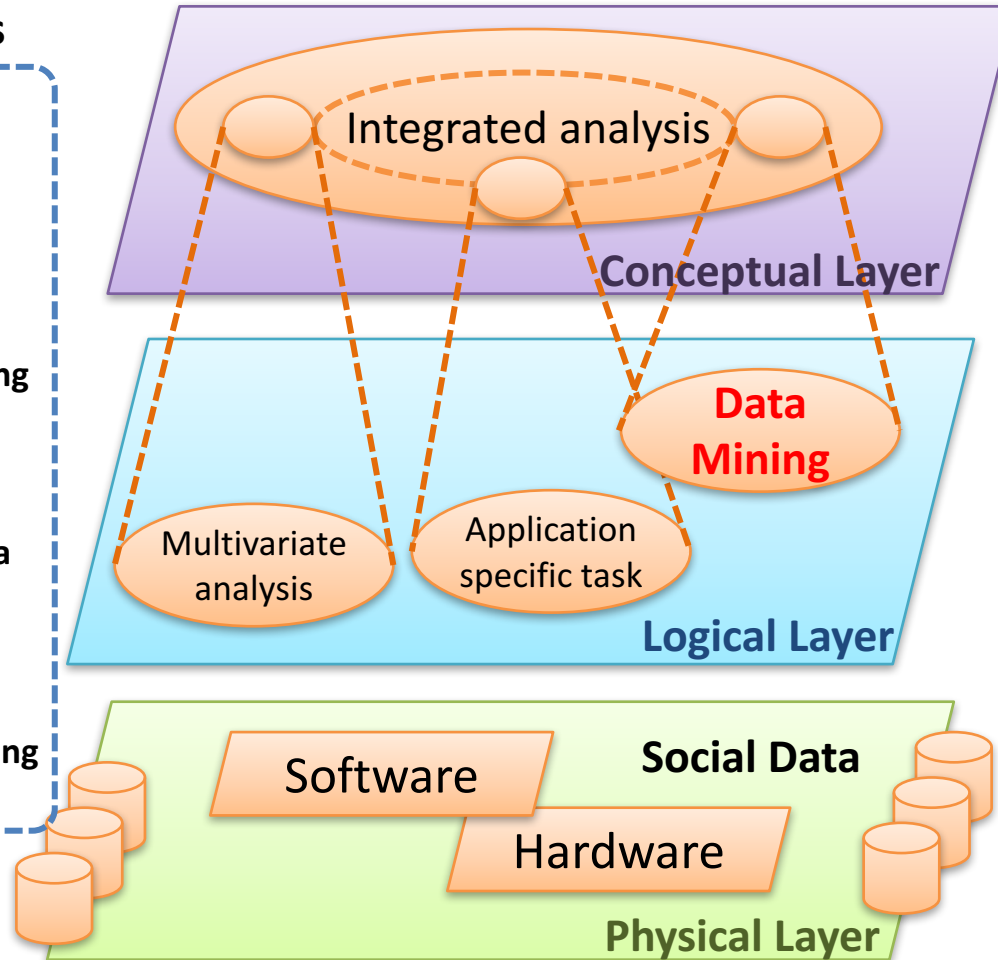


# Architecture for Social Big Data Mining

(Hiroshi Ishikawa, 2015)

## Enabling Technologies

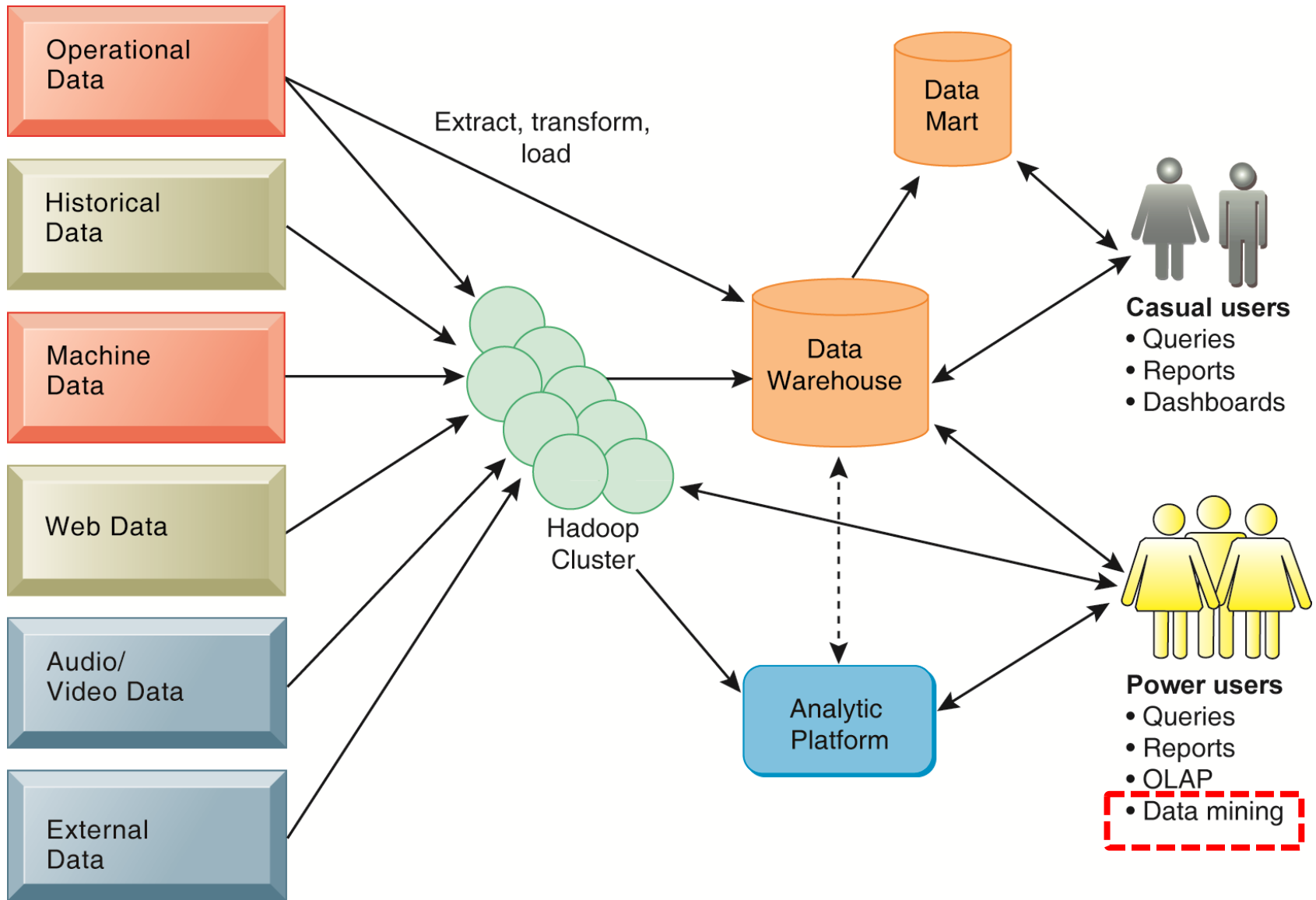
- Integrated analysis model
- Natural Language Processing
- Information Extraction
- Anomaly Detection
- Discovery of relationships among heterogeneous data
- Large-scale visualization
- Parallel distributed processing



## Analysts

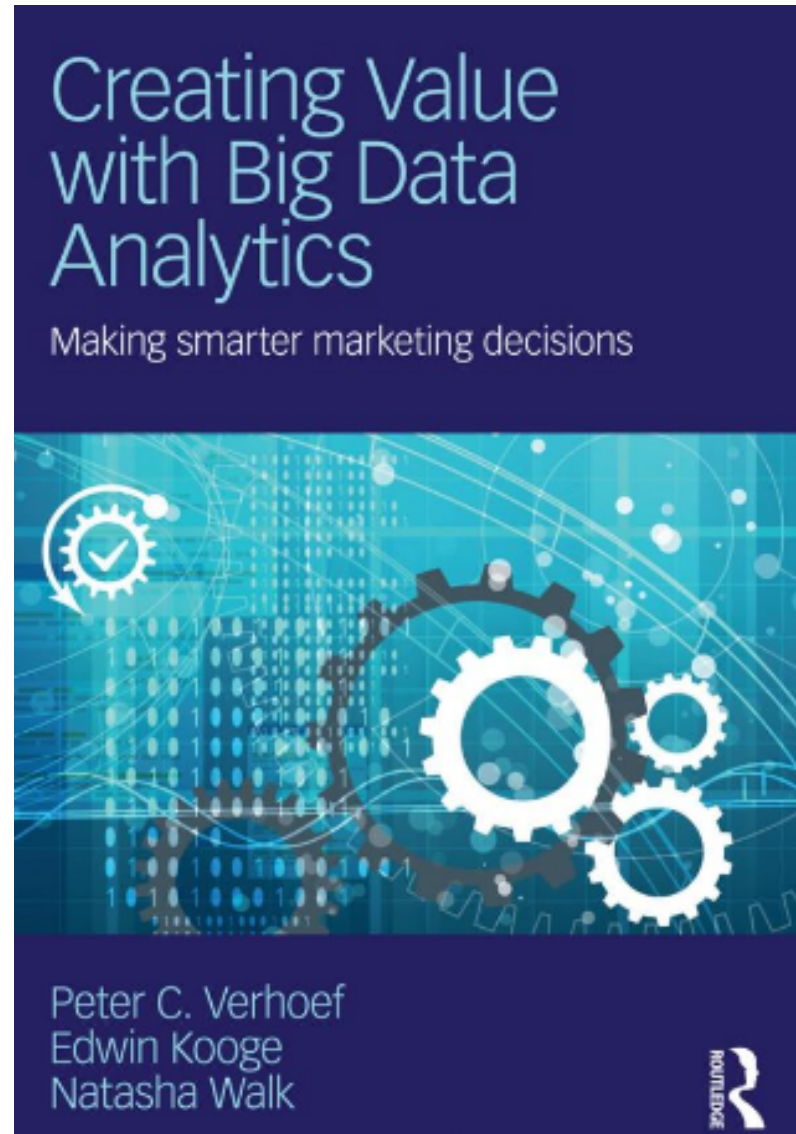
- Model Construction
- Explanation by Model
- Construction and confirmation of individual hypothesis
- Description and execution of application-specific task

# Business Intelligence (BI) Infrastructure



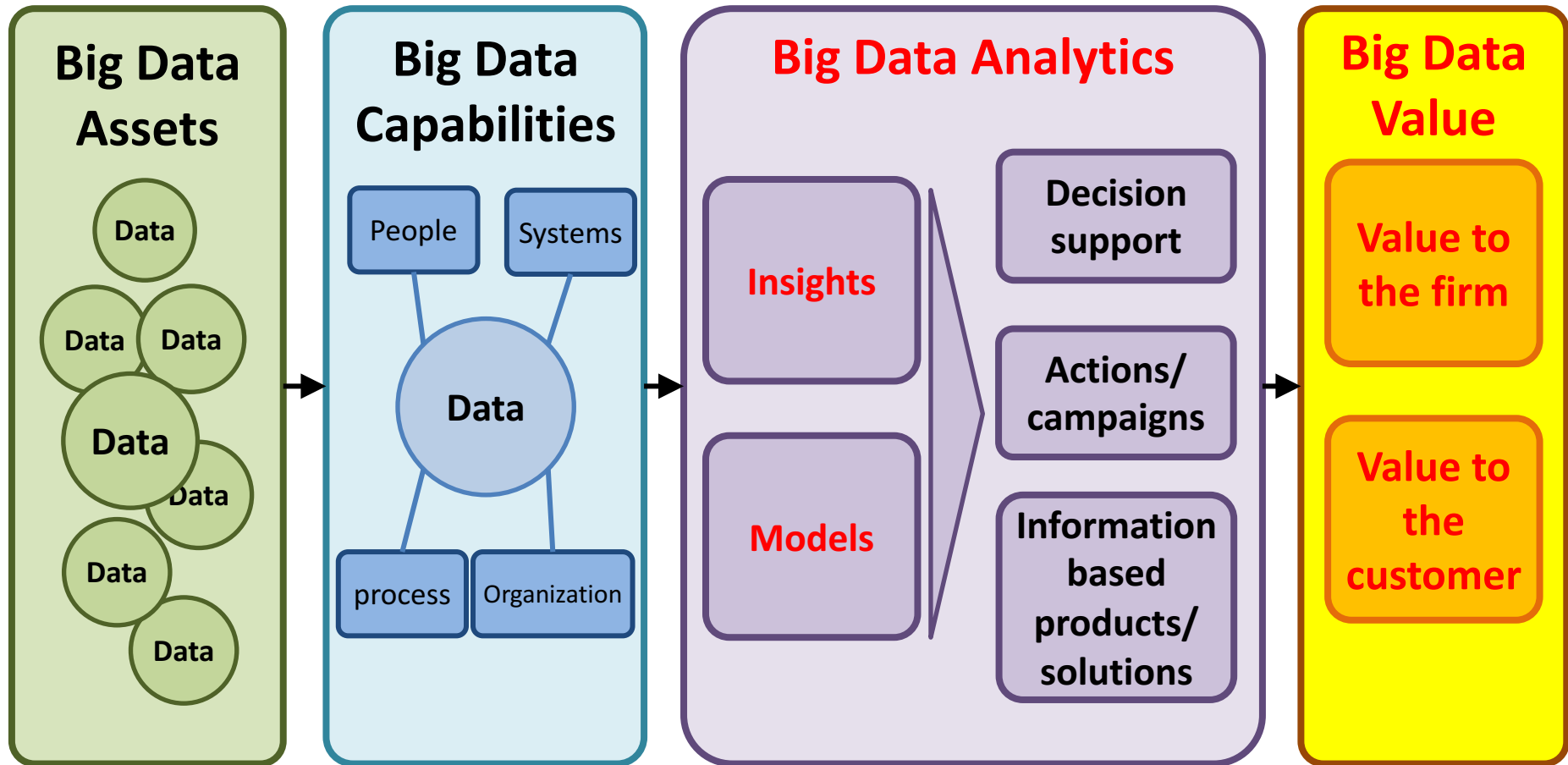


Creating Value with Big Data Analytics:  
Making Smarter Marketing Decisions,  
Peter C. Verhoef and Edwin Kooge, Routledge, 2016



# Big Data Value Creation Model

Creating Value with Big Data Analytics:  
Making Smarter Marketing Decisions



# Digital Data Platform for Enterprises

## Big Data Analytics

### Enterprise Applications



Operational  
Benchmark

Customer  
focus

Organization  
Connections

Document  
Search

Sales  
Forecast

Security (Authentication, Authorization, Auditing, Encryption, Protection)

Variety of  
Sources



Ingestion  
layer

Data  
Connectors

Data Extraction

CDC

Data Quality

Processing  
Layer



Data Mining

Data  
Enrichment

Real-time  
Streaming

Batch  
Processing

Storage  
Layer



Hadoop

NoSQL

RDBMS

In-Memory

Analytics  
Layer



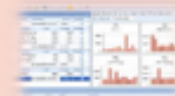
Traditional  
Analytics

Search Based  
Analytics

Predictive  
Analytics

Ad-hoc  
Analytics

Visualization  
Apps

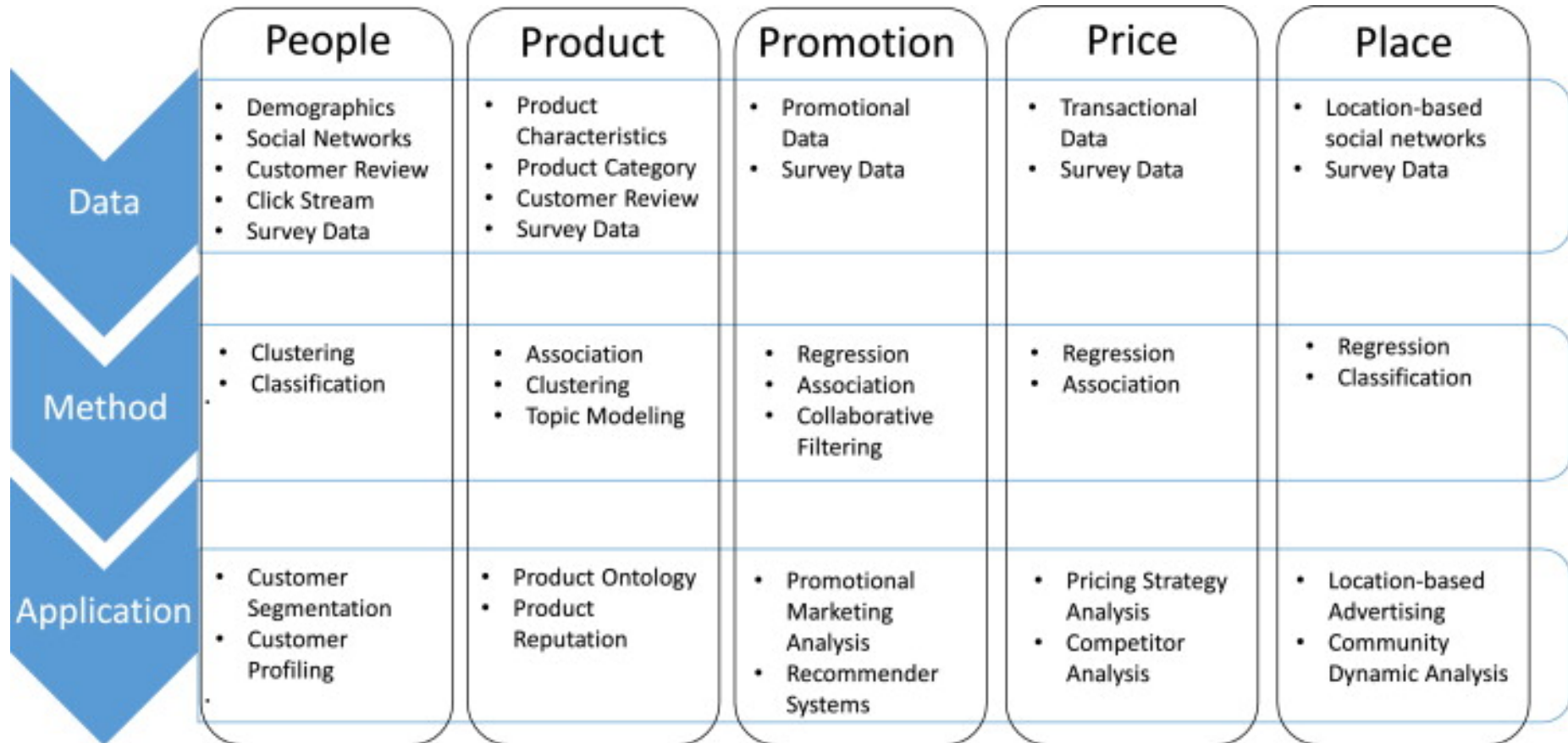


Data Governance and Monitoring (Workflow, lifecycle management, scheduler, manage)

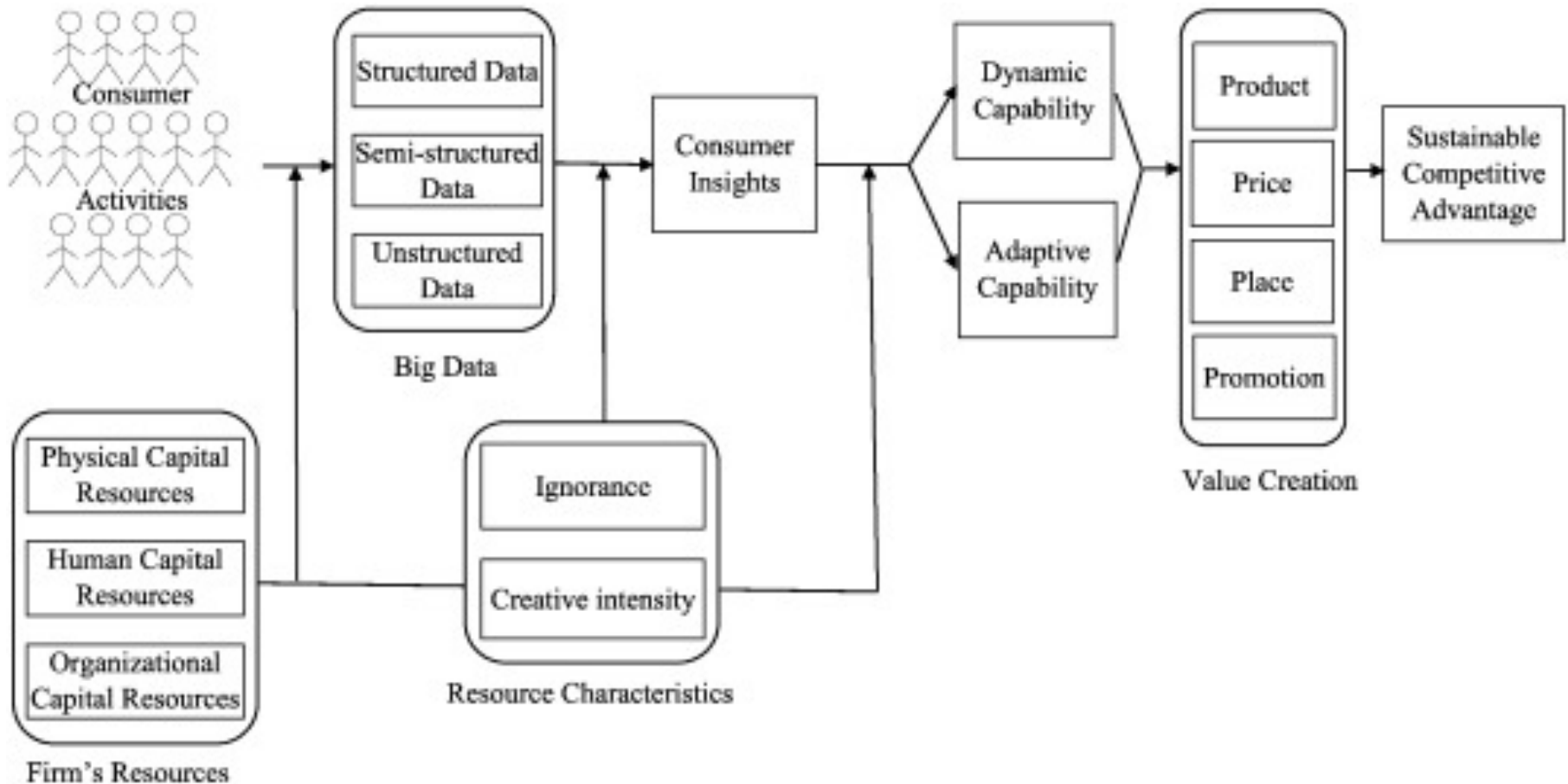


Digital Data Driven Platform for Enterprises

# A Marketing Mix Framework for Big Data Management



# A resource-based view of the impact of Big Data on competitive advantage



**LeCun, Yann,  
Yoshua Bengio,  
and Geoffrey Hinton.**

**"Deep learning."**

**Nature 521, no. 7553 (2015): 436-  
444.**

## Deep learning

Yann LeCun<sup>1,2</sup>, Yoshua Bengio<sup>3</sup> & Geoffrey Hinton<sup>4,5</sup>

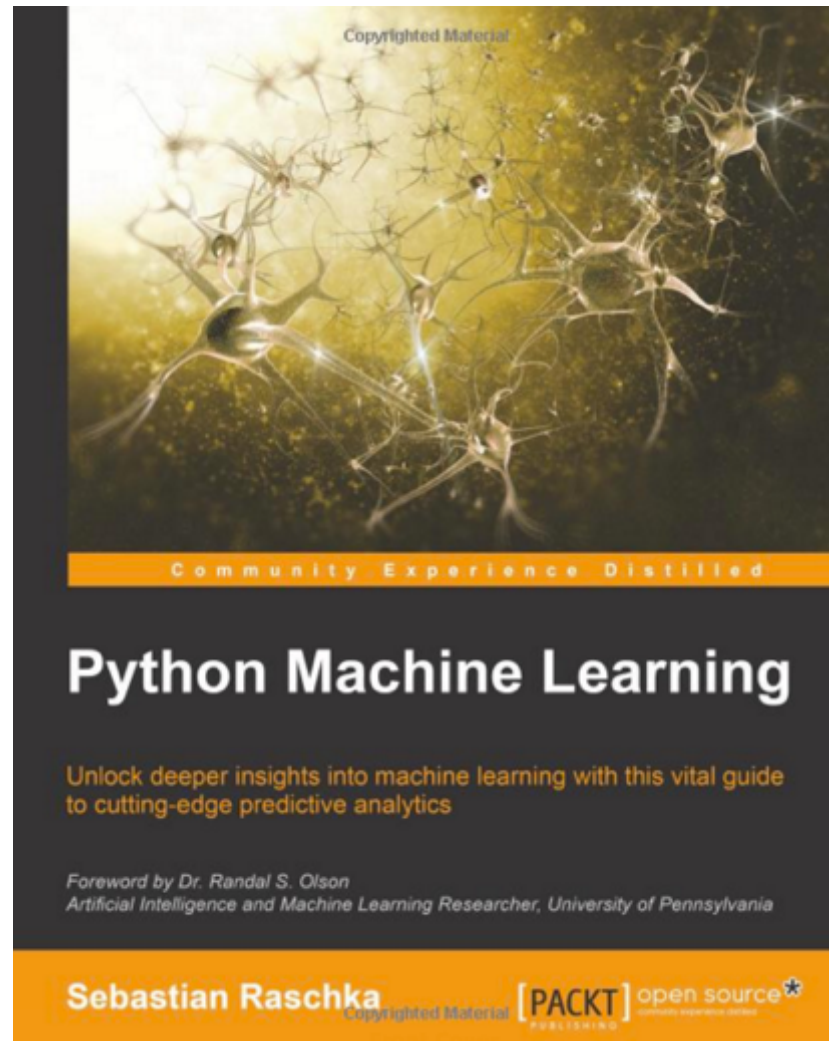
**Deep learning allows computational models that are composed of multiple processing layers to learn representations of data with multiple levels of abstraction. These methods have dramatically improved the state-of-the-art in speech recognition, visual object recognition, object detection and many other domains such as drug discovery and genomics. Deep learning discovers intricate structure in large data sets by using the backpropagation algorithm to indicate how a machine should change its internal parameters that are used to compute the representation in each layer from the representation in the previous layer. Deep convolutional nets have brought about breakthroughs in processing images, video, speech and audio, whereas recurrent nets have shone light on sequential data such as text and speech.**

**M**achine-learning technology powers many aspects of modern society: from web searches to content filtering on social networks to recommendations on e-commerce websites, and it is increasingly present in consumer products such as cameras and smartphones. Machine-learning systems are used to identify objects in images, transcribe speech into text, match news items, posts or products with users' interests, and select relevant results of search. Increasingly, these applications make use of a class of techniques called deep learning.

Conventional machine-learning techniques were limited in their ability to process natural data in their raw form. For decades, con-

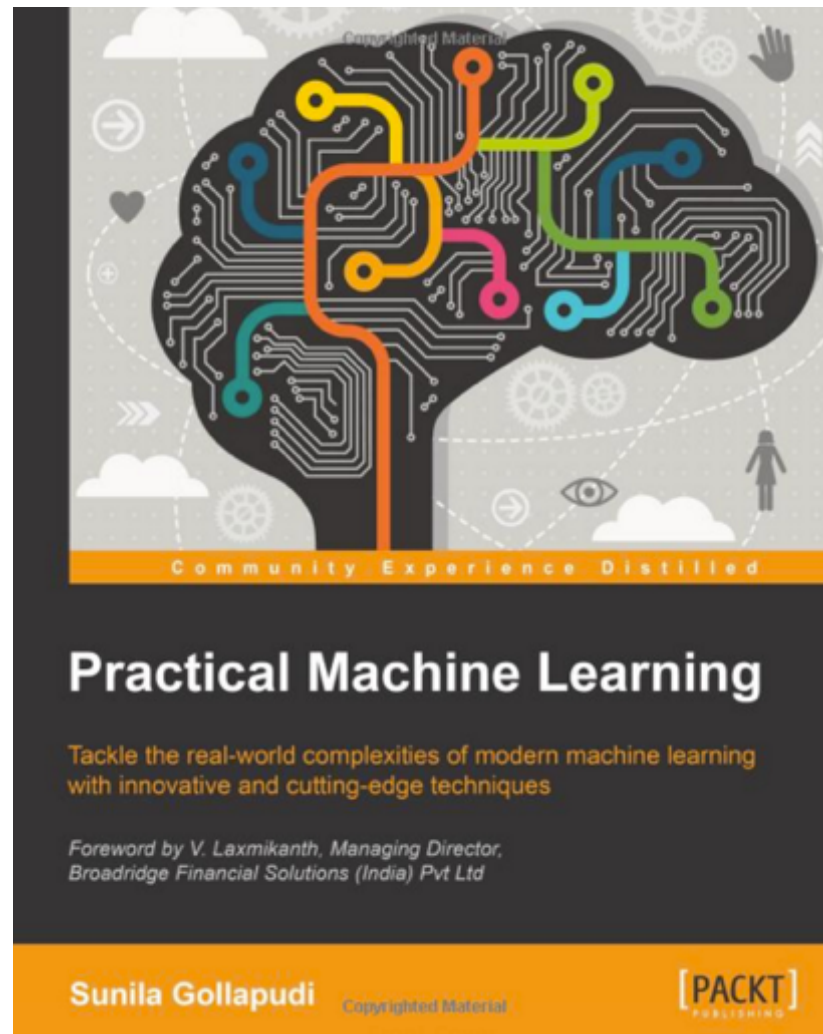
intricate structures in high-dimensional data and is therefore applicable to many domains of science, business and government. In addition to beating records in image recognition<sup>1-4</sup> and speech recognition<sup>5-7</sup>, it has beaten other machine-learning techniques at predicting the activity of potential drug molecules<sup>8</sup>, analysing particle accelerator data<sup>9,10</sup>, reconstructing brain circuits<sup>11</sup>, and predicting the effects of mutations in non-coding DNA on gene expression and disease<sup>12,13</sup>. Perhaps more surprisingly, deep learning has produced extremely promising results for various tasks in natural language understanding<sup>14</sup>, particularly topic classification, sentiment analysis, question answering<sup>15</sup> and language translation<sup>16,17</sup>.

Sebastian Raschka (2015),  
**Python Machine Learning,**  
Packt Publishing





Sunila Gollapudi (2016),  
**Practical Machine Learning,**  
Packt Publishing



# Machine Learning Models

Deep Learning

Association rules

Decision tree

Clustering

Bayesian

Kernel

Ensemble

Dimensionality reduction

Regression Analysis

Instance based

# Data Scientist

# 資料科學家



# Deep Learning

## Intelligence from Big Data



# **Data Scientist:** **The Sexiest Job** **of the 21st Century**

**(Davenport & Patil, 2012)(HBR)**

# Data Scientist:

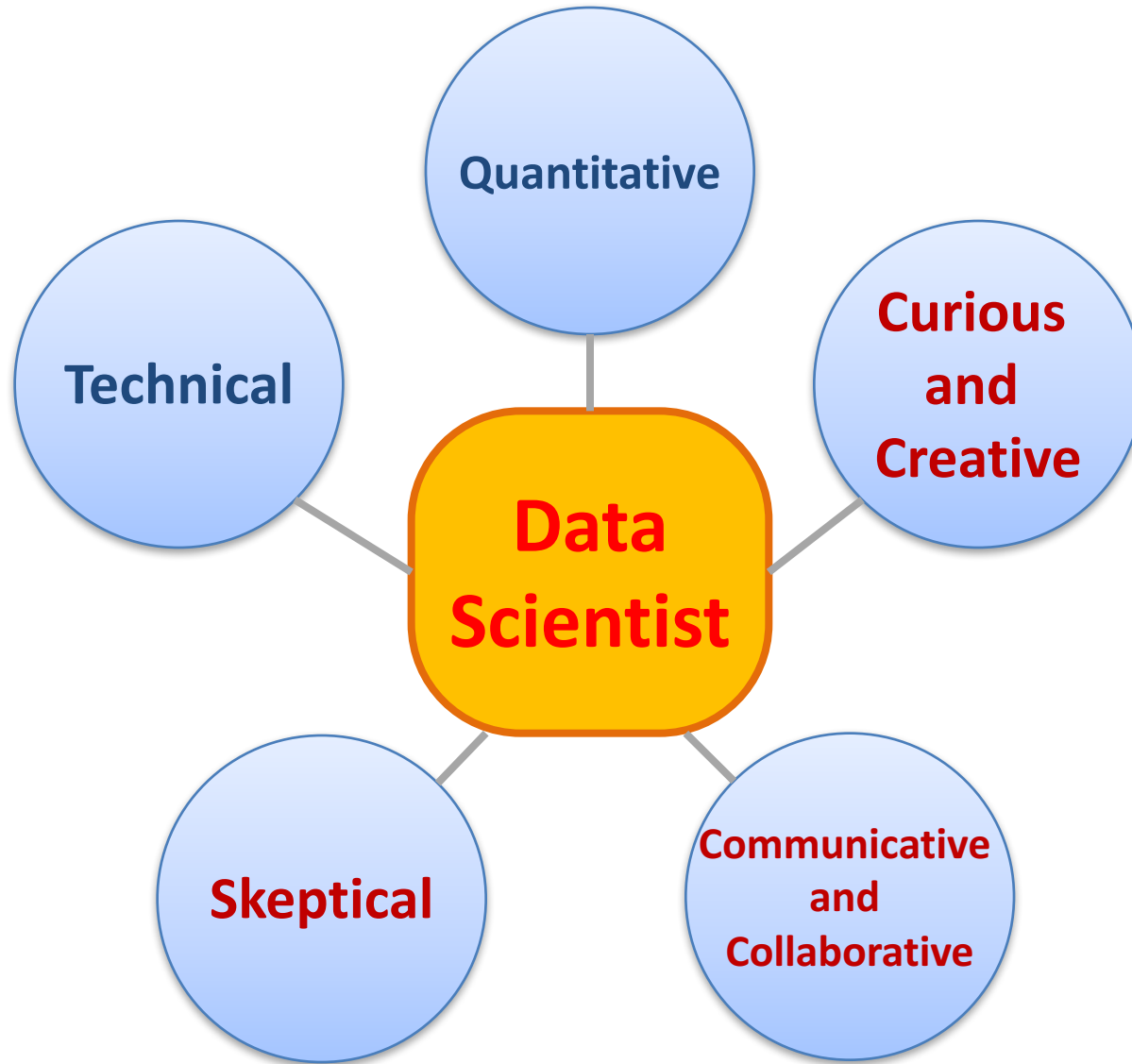
## *The Sexiest Job of the 21st Century*

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

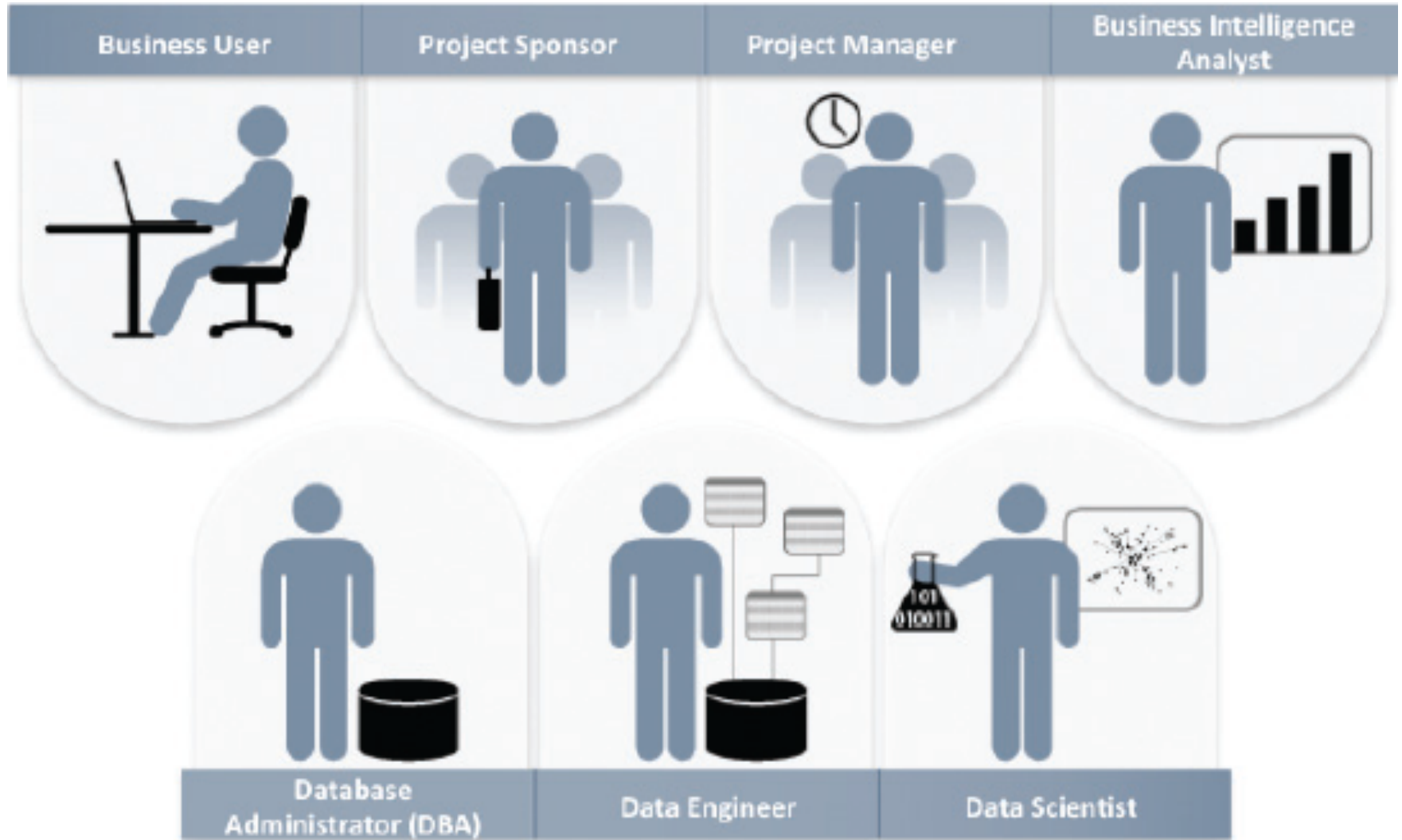
*by Thomas H. Davenport  
and D.J. Patil*

**W**hen 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."

# Data Scientist Profile

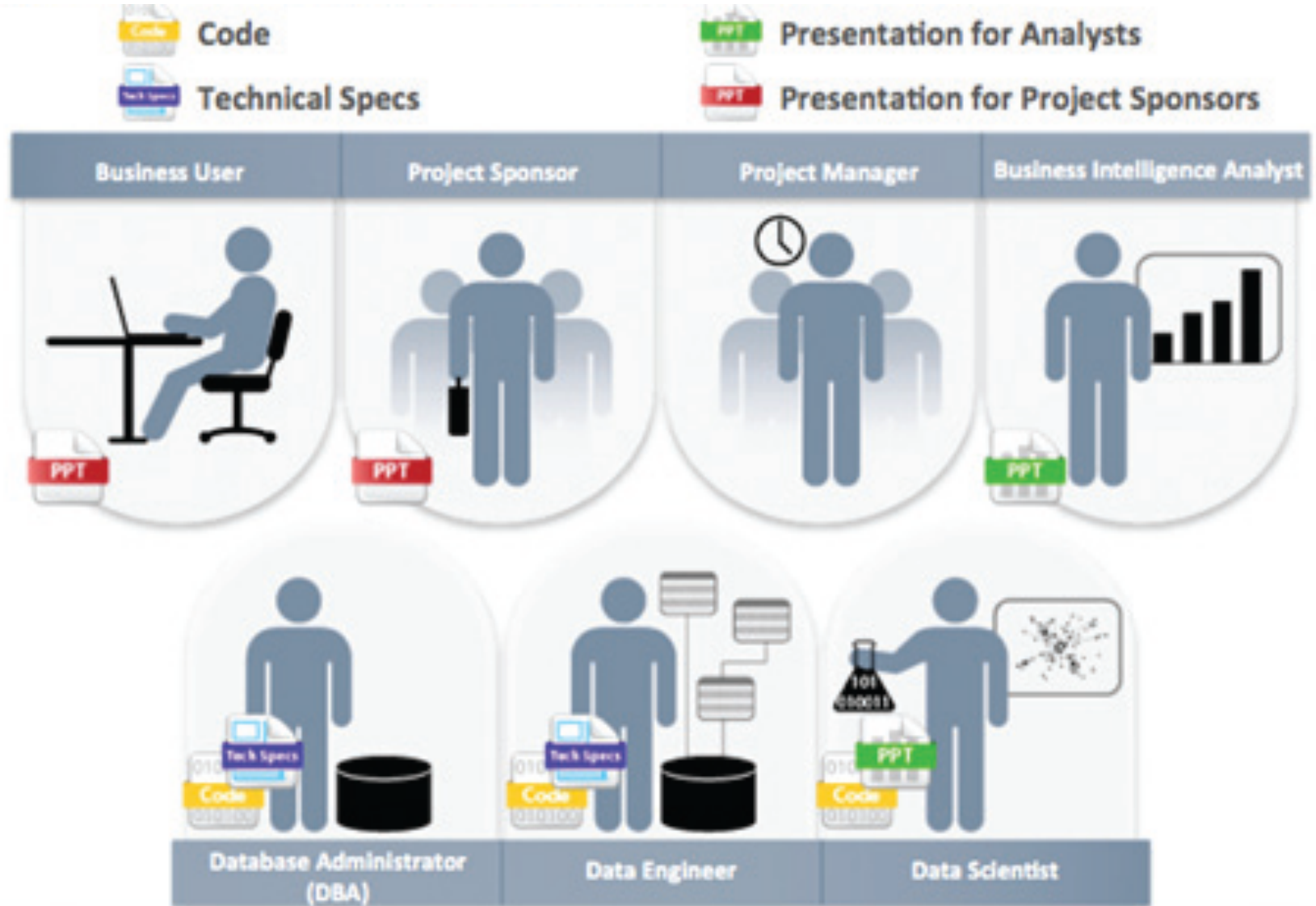


# Key Roles for a Successful Analytics Project





# Key Outputs from a Successful Analytics Project



# Data Science vs. Big Data vs. Data Analytics

## Data Science **VS** Big Data **VS** Data Analytics

DATA IS GROWING FASTER THAN EVER BEFORE.



Each person-  
**1.7 megabytes**  
created



# Data Science vs. Big Data vs. Data Analytics

## WHAT ARE THEY?



**Data Science** is a field that comprises of everything that related to data cleansing, preparation, and analysis.



**Big Data** is something that can be used to analyze insights which can lead to better decision and strategic business moves.



**Data Analytics** Involves automating insights into a certain dataset as well as supposes the usage of queries and data aggregation procedures.

# What are they used?

Data Science algorithms are used in industries like:



Big Data is used in industries like:



Data Analytics is used in industries like:



# Data Science

## What are the Skills Required?



### DATA SCIENTIST

- In-depth knowledge in SAS and/or R
- Python coding
- Hadoop platform
- SQL database/coding
- Working with unstructured data

### BIG DATA SPECIALIST

- Analytical skills
- Creativity
- Mathematics and
- Statistical skills
- Computer science
- Business skills

### DATA ANALYST

- Programming skills
- Statistical skills
- Mathematics
- Machine learning skills
- Data wrangling skills
- Communication and Data Visualization skills
- Data Intuition

**DATA SCIENTIST**

**\$113,436**  
per year.

**BIG DATA SPECIALIST**

**\$62,066**  
per year.

**DATA ANALYST**

**\$60,476**  
per year.

# Big Data Landscape 2016

## Infrastructure

**Hadoop On-Premise**  
 cloudera Hortonworks  
 MMAPR Pivotal  
 IBM InfoSphere  
 splunk jethro

**Hadoop in the Cloud**  
 amazon web services Google Cloud Platform  
 Microsoft Azure IBM InfoSphere  
 CAZENA altilscale  
 Quale xplenty

**Spark**  
 databricks  
 GridGain  
 TACHYON NEXUS

**Cluster Services**  
 amazon web services  
 Kubernetes  
 HPCC SYSTEMS  
 MESOSPHERE  
 Core OS pepperdata  
 StackIQ

## Analytics

**Analyst Platforms**  
 Palantir  
 AYASDI  
 Quid enigma  
 Digital Reasoning  
 ORBITAL INSIGHT

**Analytics Platforms**  
 Microsoft  
 guavus  
 Datameer  
 interana

**Data Science Platforms**  
 context relevant  
 CONTINUUM DataRobot  
 Alpine ADATAO  
 MODE ploity  
 dataiku Ionian  
 DOMINO sense  
 yhat ALGORITHMIA

**Visualization**  
 +ableau  
 Google Cloud Platform  
 Roambi  
 Qlik  
 CHARTIO

## Applications

**Sales & Marketing**  
 RADIUS Gainsight  
 bloomreach Zeta  
 livefyre blueyonder  
 kahuna Lattice  
 persado infer sense  
 AVISO ACTIONIQ  
 QUANTIFIND ENGA GIO

**Customer Service**  
 MEDALLIA  
 ATTENITY CLARABRIDGE  
 STELLAService  
 NGDATA Preact  
 DigitalGenius wiseia  
 appurri  
 fuse machines

**Human Capital**  
 gild  
 Connectifier  
 textio  
 entelo  
 hiQ

**Legal**  
 RAVEL  
 JUDICATA  
 Everlaw  
 Brevia  
 PREMIONION

**NoSQL Databases**  
 amazon DynamoDB Google Cloud Platform  
 Microsoft Azure ORACLE  
 mongoDB MarkLogic  
 DATASTAX  
 KEROPIKE Couchbase  
 SequoiaDB redislabs influxdata

**NewSQL Databases**  
 SAP HANA Clustrix Pivotal  
 paradigm4  
 memsql nuODB  
 MariaDB VOLTDB citusdata  
 deopdb Trafodion Cockroach LABS

**BI Platforms**  
 Power BI amazon web services  
 DOMO  
 Wave Analytics  
 GoodData birst  
 kyvos insights  
 platfora looker  
 atscale ARCADIA  
 SIBSENSE

**Statistical Computing**  
 SAS  
 SPSS  
 MATLAB

**Log Analytics**  
 splunk  
 sumologic  
 kibana  
 CLOUD PHYSICS  
 loggly

**Social Analytics**  
 NETBASE  
 DATASIFT  
 tracx bitly  
 syntheso  
 bottlenose  
 simplereach

**Ad Optimization**  
 MediaMath Integral  
 Ad Science  
 rocketfuel  
 OpenX theTradeDesk  
 Adgorithms  
 Liventent dstillery  
 DataXu Appier TAFAD

**Security**  
 CYLANCE  
 CounterTack cyberason  
 ThreatMetrix  
 AREA 1 SECURITY SentinelOne  
 Recorded Future Guardian Analytics  
 FORTSCALE siftscience  
 Kaybase feedzai SIGNIFYD

**Vertical AI Applications**  
 facebook  
 Clara  
 KASIST  
 lumiata

**Graph Databases**  
 neo4j  
 OrientDB  
 InfiniteGraph

**MPP Databases**  
 TERADATA  
 VERTICA  
 NETEZZA  
 kognitio  
 dremio

**Cloud EDW**  
 amazon web services Google Cloud Platform  
 Microsoft Azure Pivotal  
 snowflake  
 PAXATA  
 Infoworks

**Data Transformation**  
 alteryx  
 TRIFACTA  
 tamer  
 StreamSets  
 Alation

**Data Integration**  
 informatica  
 Put potential to work:  
 MuleSoft  
 snapLogic  
 BedrockData

**Real-Time**  
 amazon web services  
 METAMARKETS  
 confluent  
 DATATORRENT  
 dataArtisans

**Machine Learning**  
 Azure Machine Learning  
 H2O  
 SKYTREE  
 rapidminer DATASIRI  
 deepnlp VISERIE  
 PredictionIO glowfish

**Speech & NLP**  
 NarrativeScience  
 api.ai NUANCE  
 Gridspace  
 semanticmachines  
 cortico.io  
 mindmeld  
 IDIBON yseop

**Horizontal AI**  
 IBM Watson  
 Cortana sentiment  
 VIV  
 nervana  
 Numenta  
 MetaMind clarifai  
 DEXTR  
 Cosmotic Intelligence

**Publisher Tools**  
 outbrain  
 mixpanel  
 Chartbeat  
 yieldbot  
 Yieldmo

**Govt/ Regulation**  
 Socrata  
 OPENGOV  
 FN FiscalNote  
 enigma  
 PREPOL  
 mark43  
 OpenDataSoft

**Finance**  
 affirm  
 LendingClub  
 OnDeck  
 Kreditech  
 Kabbage  
 INSIGHT  
 ZUORA Dataminr  
 Lenddo  
 KENSHO AIDYIA  
 ISENTIUM  
 Quantopian  
 sentiment

**Management / Monitoring**  
 New Relic  
 APPDYNAMICS  
 actifio  
 Numerify  
 splunk  
 DATADOG  
 Trocana Anodot

**Security**  
 TANIUM  
 illumio  
 CODE42  
 DataGravity  
 CipherCloud  
 VECTRA  
 sqrrl BlueTalon

**Storage**  
 amazon web services Google Cloud Platform  
 Microsoft Azure Pivotal  
 panasas  
 nimblestorage  
 Qumulo

**App Dev**  
 apigee  
 CASK  
 Typesafe

**Crowd-sourcing**  
 amazon mechanicalturk  
 CrowdPower  
 WorkFusion

**Search**  
 hp Autonomy ORACLE  
 ENDECA  
 EXALEAD  
 Lucidworks  
 elastic ThoughtSpot  
 MAANA swifttype

**Data Services**  
 OPERA  
 Mis Sigma  
 DATA SCIENCE  
 kaggle datascience  
 DataKind

**For Business Analysts**  
 OrigamiLogic  
 ClearStory  
 CIRRO  
 import io

**SMB / Commerce**  
 Google Analytics  
 AMPITUDE RJMetrics  
 BLUECORE  
 sumAll granify  
 Airtable  
 retention custora

**Education/ Learning**  
 KNEWTON  
 Clever  
 Declara  
 PANORAMA  
 knowTe

**Life Sciences**  
 23andMe  
 Pathway Genomics  
 XRecombine  
 KYRUS FLATIRON  
 zymergen HealthTap  
 METABIOTA ZEPHYR HEALTH ovia  
 Gingerio transcriptic Glow  
 entlic AiCure Atomwise

**Industries**  
 OP@WER eHarmony  
 RetailNext  
 STITCH FIX  
 WorkFusion  
 BLUE RIVER  
 TACHYUS  
 SwiftKey  
 Seeq FarmLogs  
 HowGood  
 select  
 NIGHT MACHINE  
 statmuse BOXEVER

## Cross-Infrastructure/Analytics

amazon web services Google Microsoft IBM SAP SAS hp Autonomy vmware talent TIBCO TERADATA ORACLE NetApp

## Open Source

**Framework**  
 hadoop HOPS  
 YARN Spark  
 MESOS TEZ  
 Flink CDAP

**Query / Data Flow**  
 SLAMDATA  
 DRILL  
 Google Cloud Dataflow

**Data Access**  
 cassandra HBASE  
 mongoDB  
 CouchDB  
 riak  
 OPENTSOB

**Coordination**  
 Apache Zookeeper  
 Apache Ambari

**Real-Time**  
 STORM Spark  
 APEX Flink  
 TACHYON druid

**Stat Tools**  
 Scala  
 Numpy  
 SciPy

**Machine Learning**  
 milib  
 Apache SINGA  
 MADlib  
 Aerosolve  
 Caffe  
 FeatureFu  
 DIMSUM  
 WEKA  
 jupyter DL4J

**Search**  
 elasticsearch  
 Solr  
 Lucene

**Security**  
 Apache Ranger  
 Zeppelin

## Data Sources & APIs

**Health**  
 Apple JAWBONE GARMIN  
 practicefusion fitbit  
 Withings VALIDIC netatmo  
 kinsa Human API

**IOT**  
 UPTAKE  
 ThingWorx  
 helium samsara  
 AUGURY estimate

**Financial & Economic Data**  
 Bloomberg DOW JONES  
 YODLEE PREMISE S&P CAPITAL IQ  
 quandl xignite CB INSIGHTS  
 mattermark estimize PLAID

**Air / Space / Sea**  
 PLANET LABS  
 WINDWARD  
 CRUISE SKYCATCH  
 Airware DroneDeploy

**Location/People/Entities**  
 GARMIN foursquare InsideView esri  
 STREETLINE  
 Connecting the Real World  
 CARTODB factual PlaceIQ  
 Crismon Hexagon placemeter BASIS Sense

**Other**  
 qualtrics  
 panjiva  
 DATA.GOV

**Incubators & Schools**  
 GA DataCamp  
 INSIGHT METIS  
 DataElite  
 The Data Incubator

© Matt Turck (@mattturck), Jim Hao (@jimrhao), & FirstMark Capital (@firstmarkcap)

FIRSTMARK

# Summary

- Social Computing
- Big Data Analysis



# References

- Hiroshi Ishikawa (2015), Social Big Data Mining, CRC Press
- Jennifer Golbeck (2013), Analyzing the Social Web, Morgan Kaufmann
- Sunila Gollapudi (2016), Practical Machine Learning, Packt Publishing
- Devangana Khokhar (2015), Gephi Cookbook, Packt Publishing
- EMC Education Services, Data Science and Big Data Analytics: Discovering, Analyzing, Visualizing and Presenting Data, Wiley, 2015
- Stephan Kudyba (2014), Big Data, Mining, and Analytics: Components of Strategic Decision Making, Auerbach Publications