

ENTRUE WORLD 2012

New Frontier of the Smart World: Advanced Analytics



빅 데이터 환경의 고급 분석 기법과 지원 기술 동향

이명진 박사

연세대학교 지식정보화연구소



연세대학교



목 차

- I . 빅 데이터
- II . 고급분석 기술 동향
- III . 고급분석 지원을 위한 인프라
- IV . 향후 과제

목 차

I. 빅 데이터

II. 고급분석 기술 동향

III. 고급분석 지원을 위한 인프라

IV. 향후 과제

東

나! 데이터

넌 누구야?



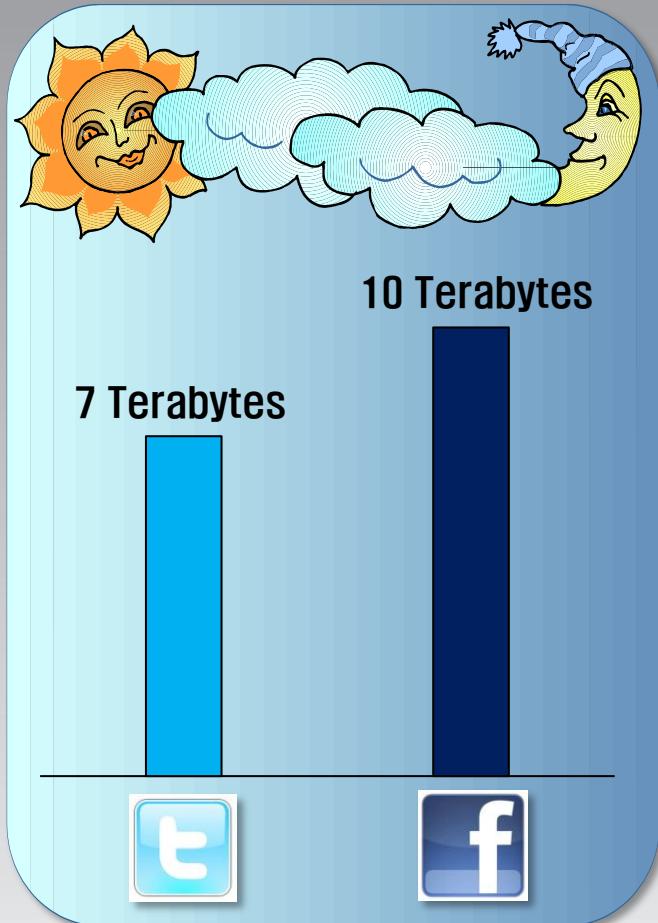
빅 데이터 시대의 도래

◀ 사용자 중심 웹 환경으로의 변화

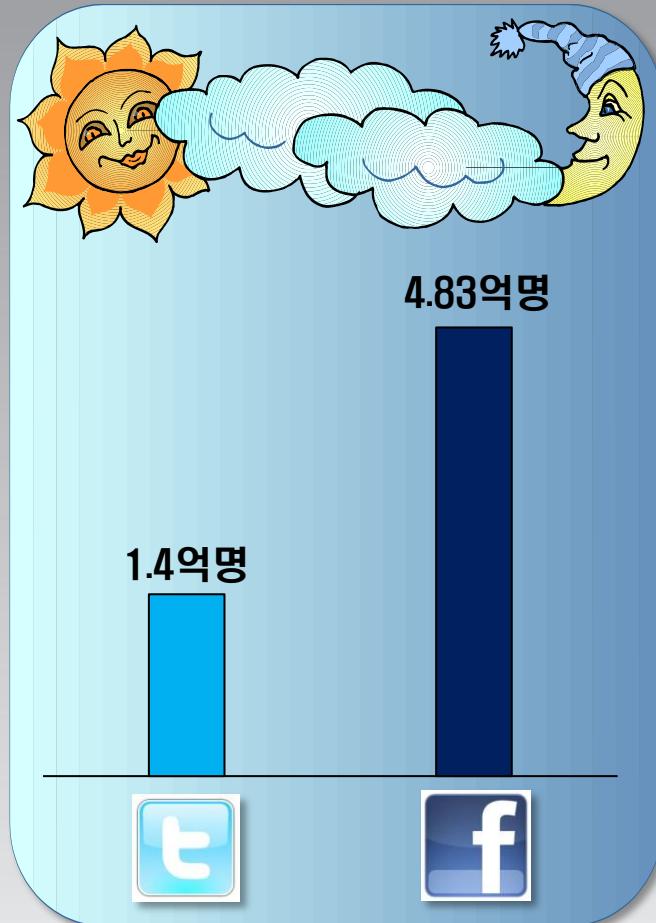


<http://iconexpo.com/2009/12/20-free-web-2-0-icons-colored-pen-version/>

빅 데이터 시대의 도래



Rod Smith, Internet Summit 2010



Leena Rao, Senior Editor in TechCrunch, Facebook

빅 데이터 시대의 도래

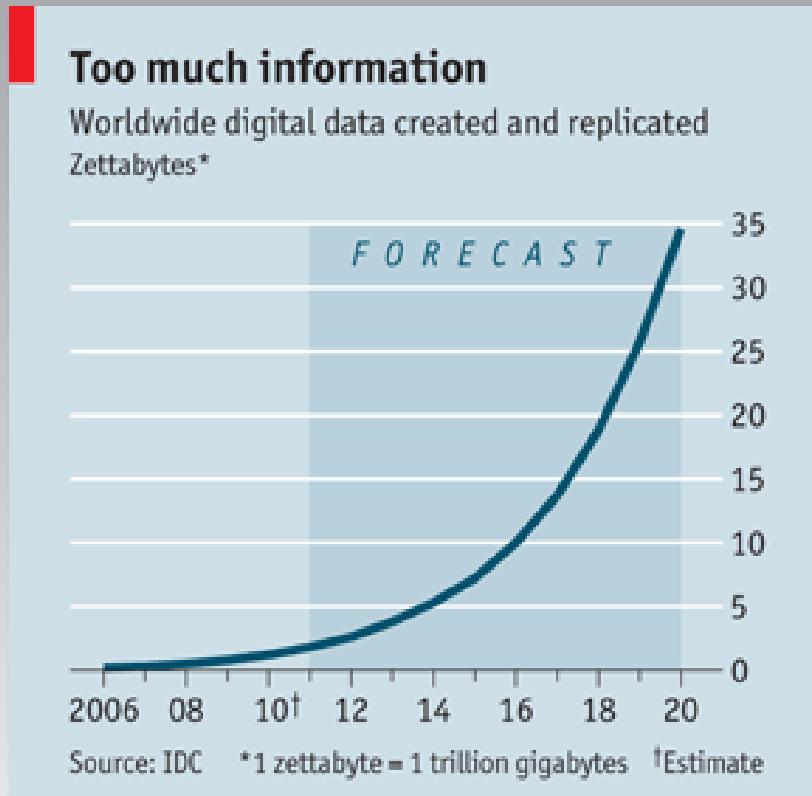
◆ 상시 인터넷 접속이 가능한 기기의 보급



마케팅인사이트, 휴대폰 기획조사

빅 데이터 시대의 도래

◆ 갈수록 증가하는 데이터의 양



빅 데이터 시대의 도래

◀ 이미 축적된 기업의 데이터

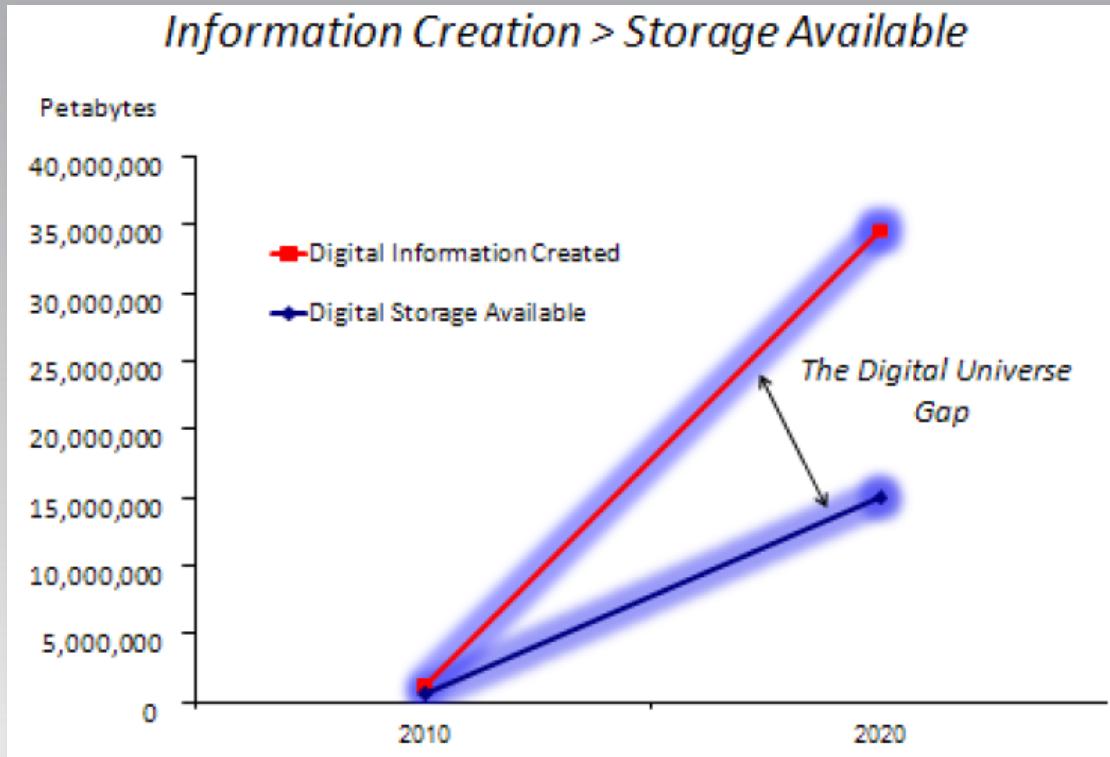


빅 데이터의 정의



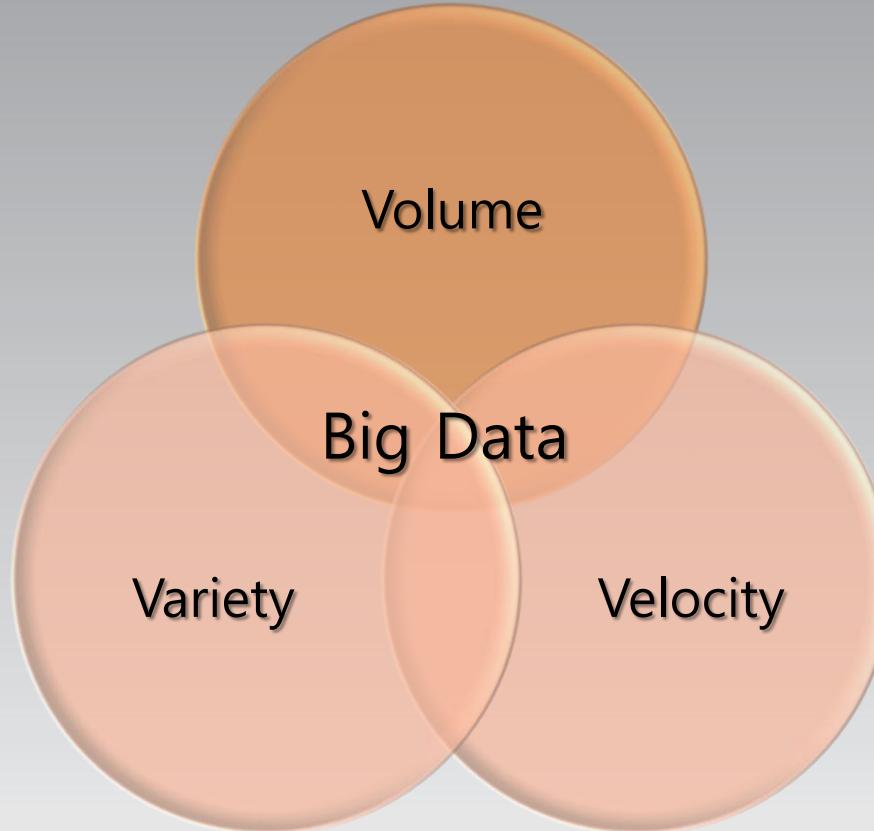
<http://tubbytoon.deviantart.com/art/Slim-Fat-Girl-91422224>

- ◆ 기존의 데이터베이스 도구의 파일화, 저장, 관리 및 분석의 역량을 넘어서는 크기의 데이터 집합 - IDC



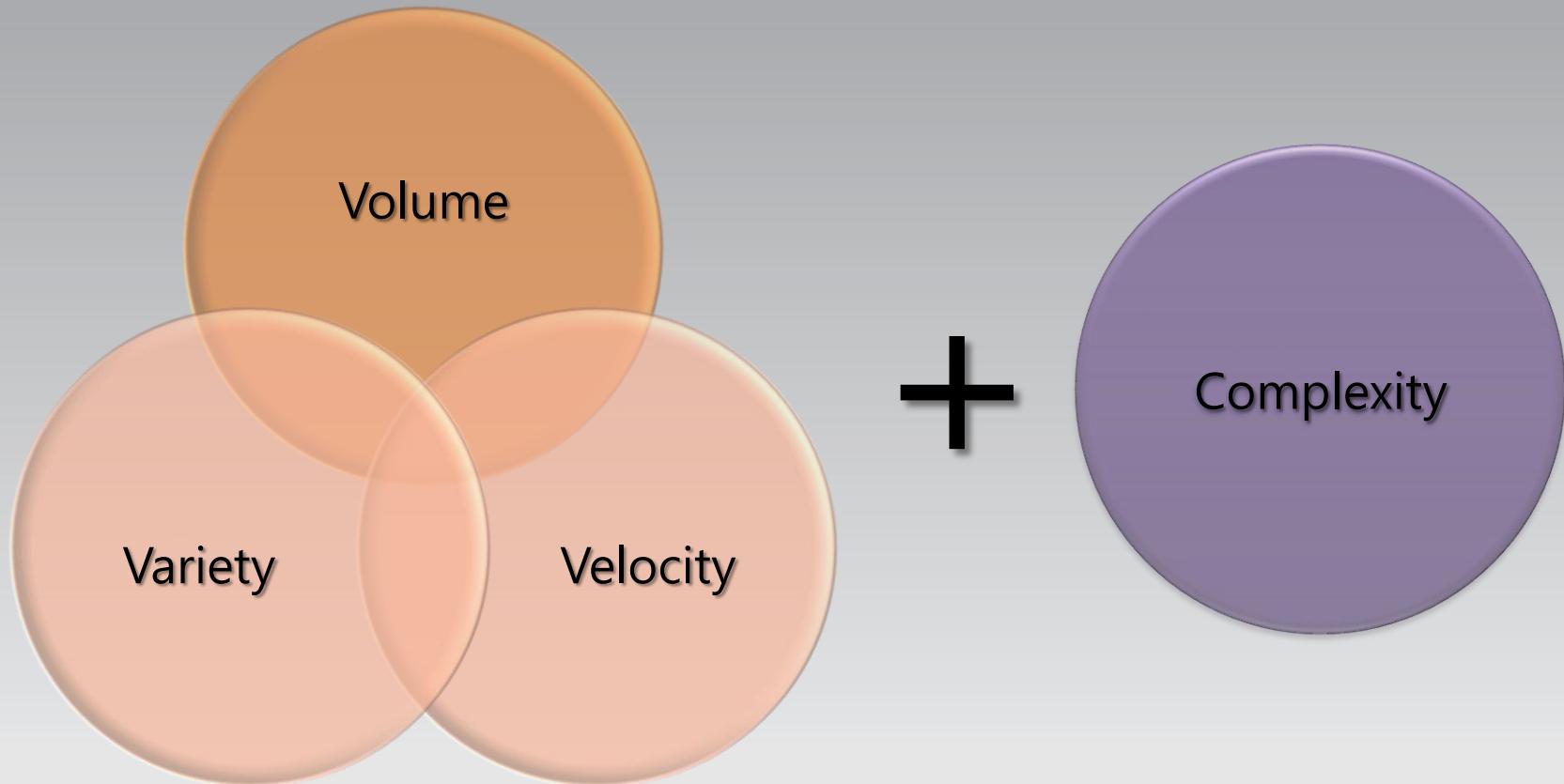
빅 데이터의 정의

- 3Vs: Volume, Velocity, Variety - IBM



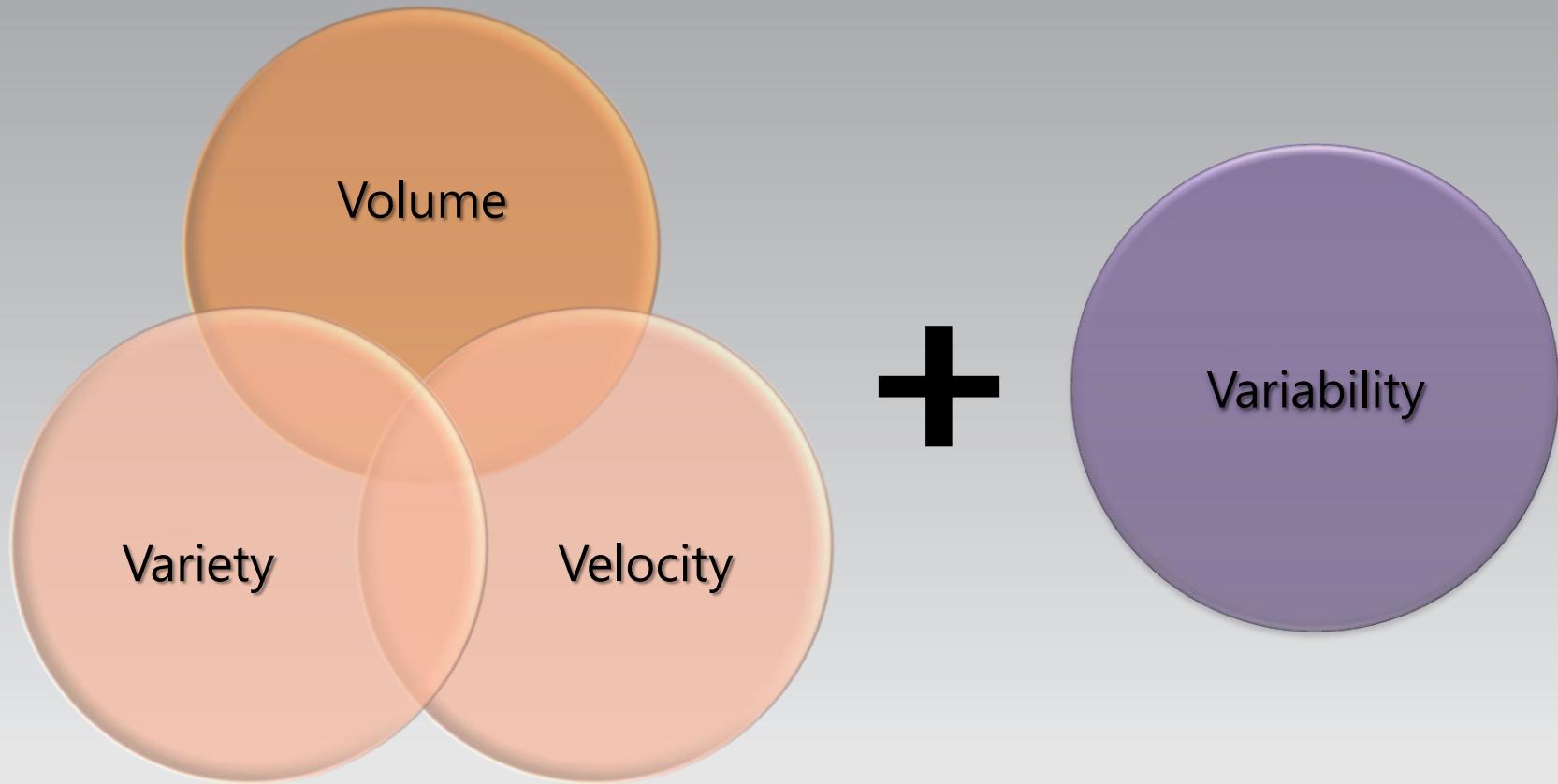
빅 데이터의 정의

- ◆ 3Vs + Complexity - Gartner



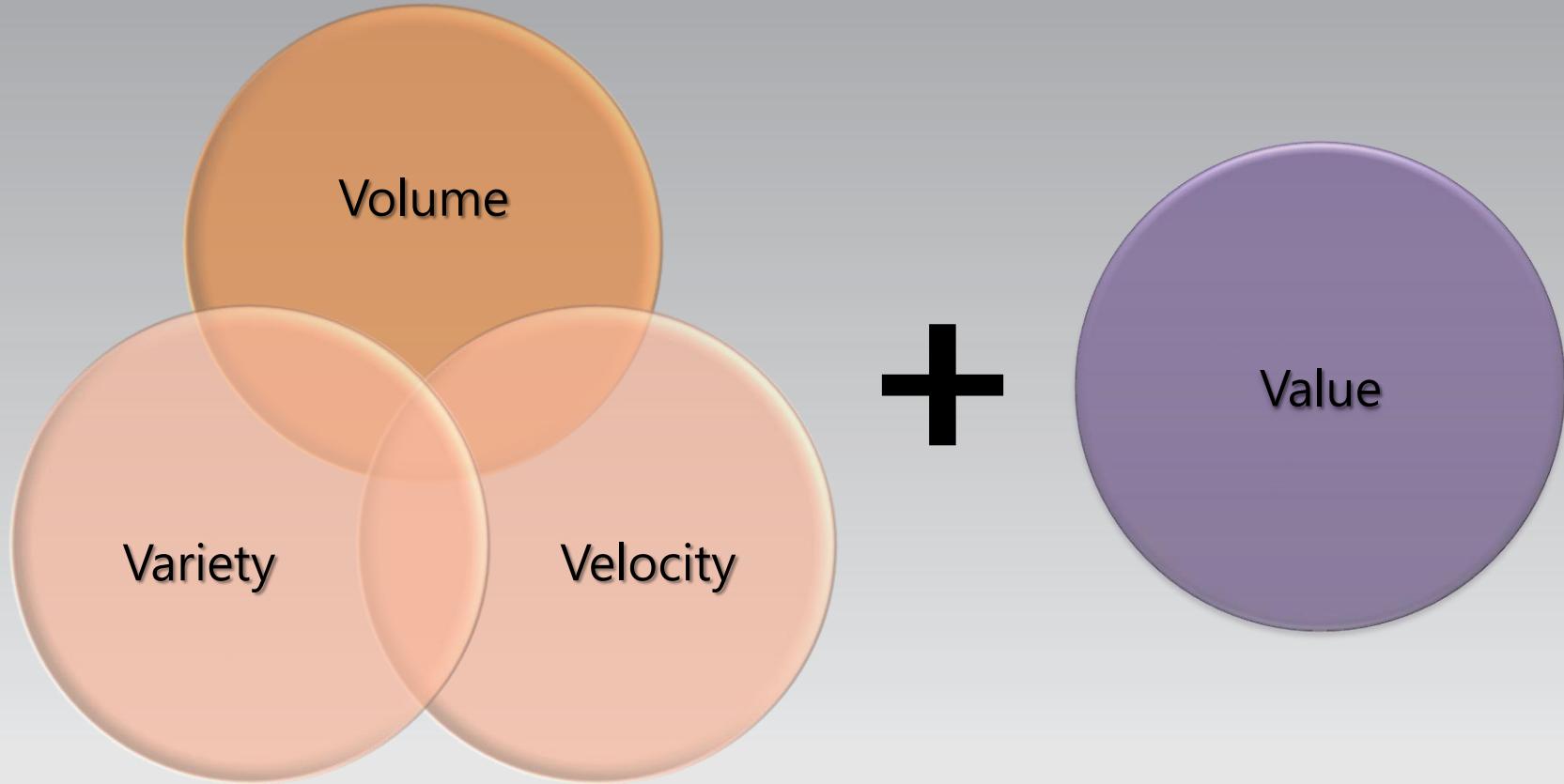
빅 데이터의 정의

- 3Vs + Variability - Forrester

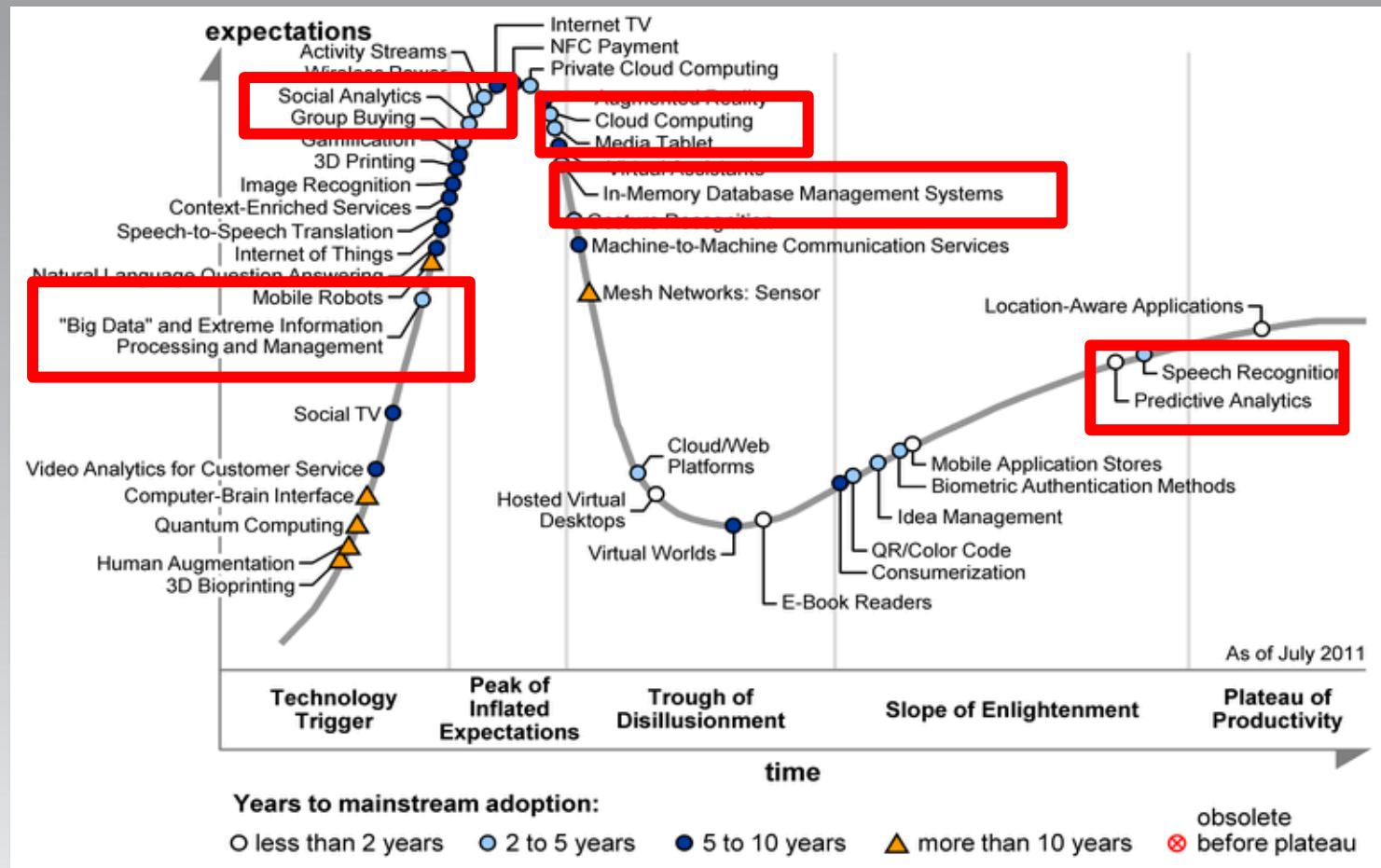


빅 데이터의 정의

- 3Vs + Value - Oracle



빅 데이터의 현재 위치



Gartner – Hype Cycle 2012

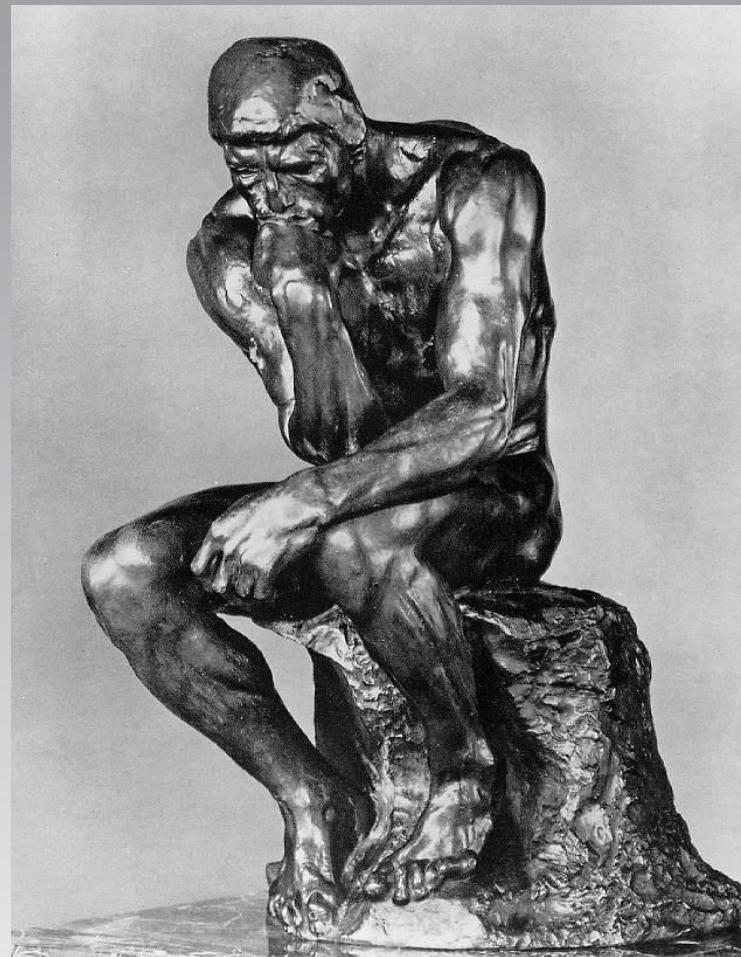
The Obama Administration this morning unveiled details about its Big Data R&D Initiative, committing more than **\$200 million in new funding** through six agencies and departments to improve “**our ability to extract knowledge and insights from large and complex collections of digital data.**”

- March 29th, 2012



생각해 볼 문제는...?

- ◆ 빅 데이터를 이용하여 비즈니스에서 적용될 수 있는 어떠한 가치를 만들어 내며, 어떻게 이를 활용할 것인가
- ◆ 빅 데이터를 기업의 부가가치를 창출할 수 있는 하나의 도구로써 활용



<http://www.amnh.org/exhibitions/brain/thinking.php>

그렇다면 우리의 역할은...?



기업의 내외부에서 생성된 데이터를
저장하며 처리 및 분석 과정을 거쳐
그 결과를 데이터의 소비자에게까지
효율적으로 보여줄 수 있는 일련의
모든 과정과 이를 지원하기 위한 기
술을 활용

<http://fireflyfoundation.wordpress.com/>

우리에게 필요한 것...



지식 수준의 정보를 발견하기 위한
분석 기술



이를 지원하기 위한
기술 인프라

목 차

I. 빅 데이터

II. 고급분석 기술 동향

III. 고급분석 지원을 위한 인프라

IV. 향후 과제

고급 분석이 필요한 이유

- ◆ 고객 및 사용자의 선호도와 그들의 요구사항을 빠르고 효과적으로 반영할 수 있는 의사결정 기술이 기업이 경쟁우위를 차지하는 주요 요소로 작용

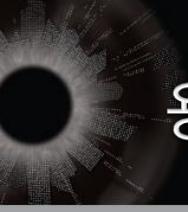
- Ranjit Bose, University of New Mexico

- ◆ 빅 데이터 환경에서 기업의 경쟁우위를 선점하기 위한 의사결정을 위해 기업의 비즈니스 가치 창출을 할 수 있는 분석 기술이 필요

고급 분석의 중요성

2010	2011	2012
Cloud Computing	Cloud Computing	Media Tablets and Beyond
Advanced Analytics	Mobile Applications and Media Tablets	Mobile-Centric Applications and Interfaces
Client Computing	Social Communications and Collaboration	Contextual and Social User Experience
IT for Green	Next Generation Analytics	Internet of Things
Reshaping the Data Center	Video	App Stores and Marketplaces
Social Computing	Social Analytics	Next-Generation Analytics
Security – Activity Monitoring	Context-Aware Computing	Big Data
Flash Memory	Storage Class Memory	In-Memory Computing
Virtualization for Availability	Ubiquitous Computing	Extreme Low-Energy Servers
Mobile Applications	Fabric-Based Infrastructure and Computers	Cloud Computing

Gartner - the Top 10 Strategic Technologies



용어 정리부터...

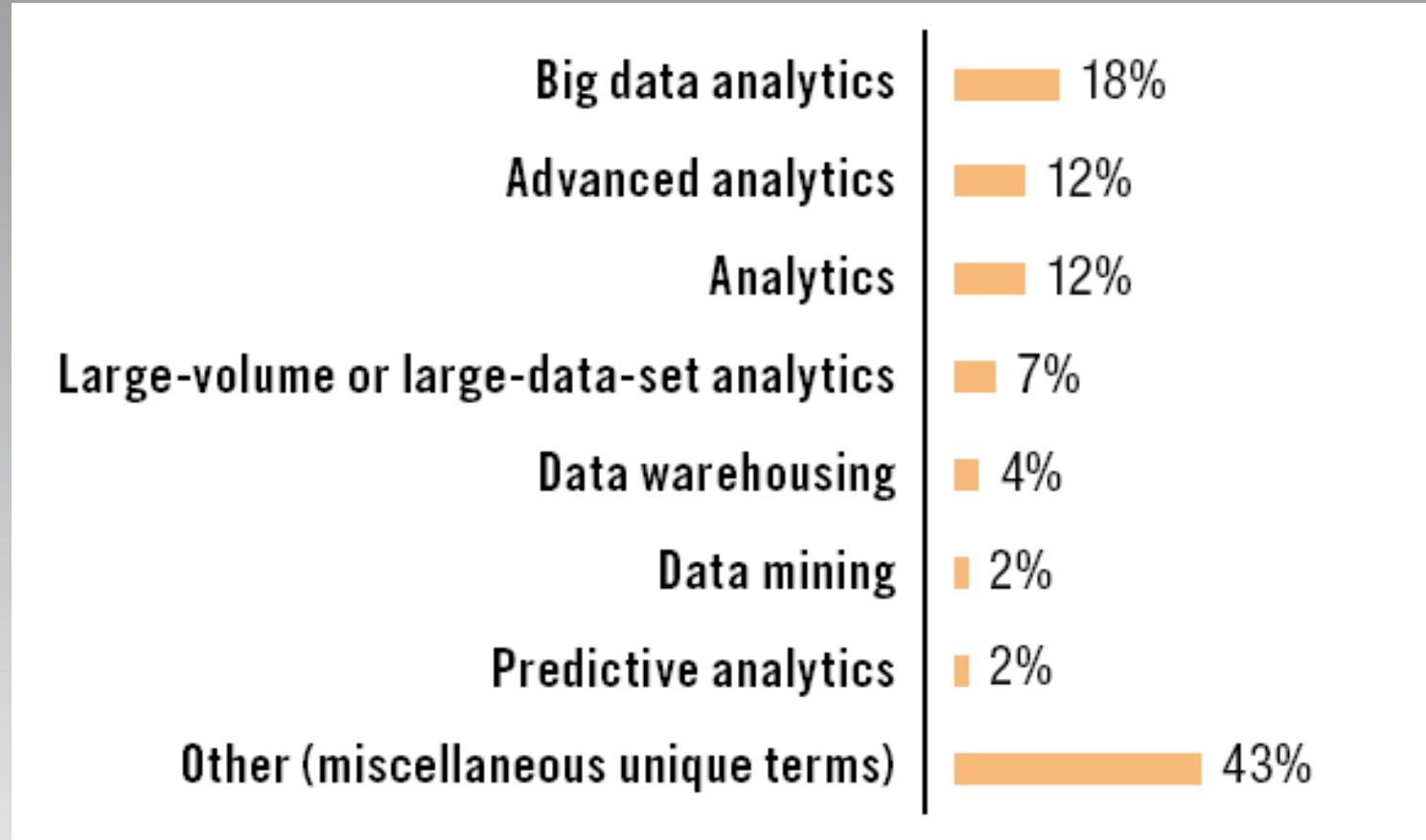


빅 데이터 분석 (Big Data Analytics)?

고급분석 (Advanced Analytics)?

BI에서의 분석 (Analytics)?

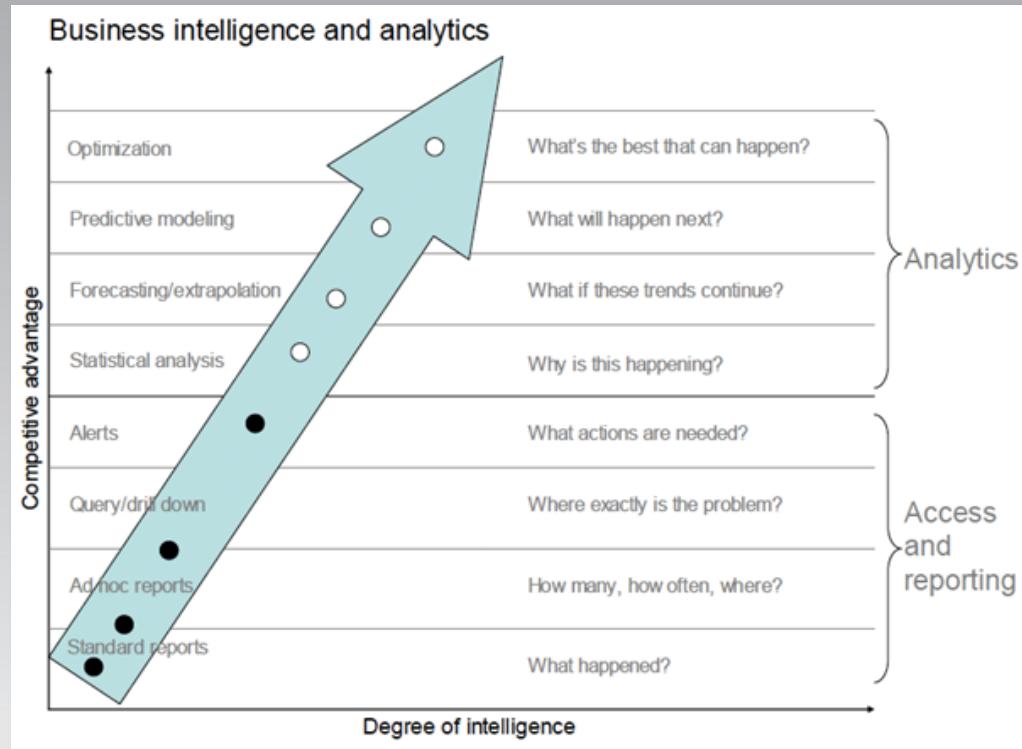
빅 데이터 분석과 고급분석



Philip Russom, Director of TDWI Research

비즈니스 인텔리전스에서의 비즈니스 분석

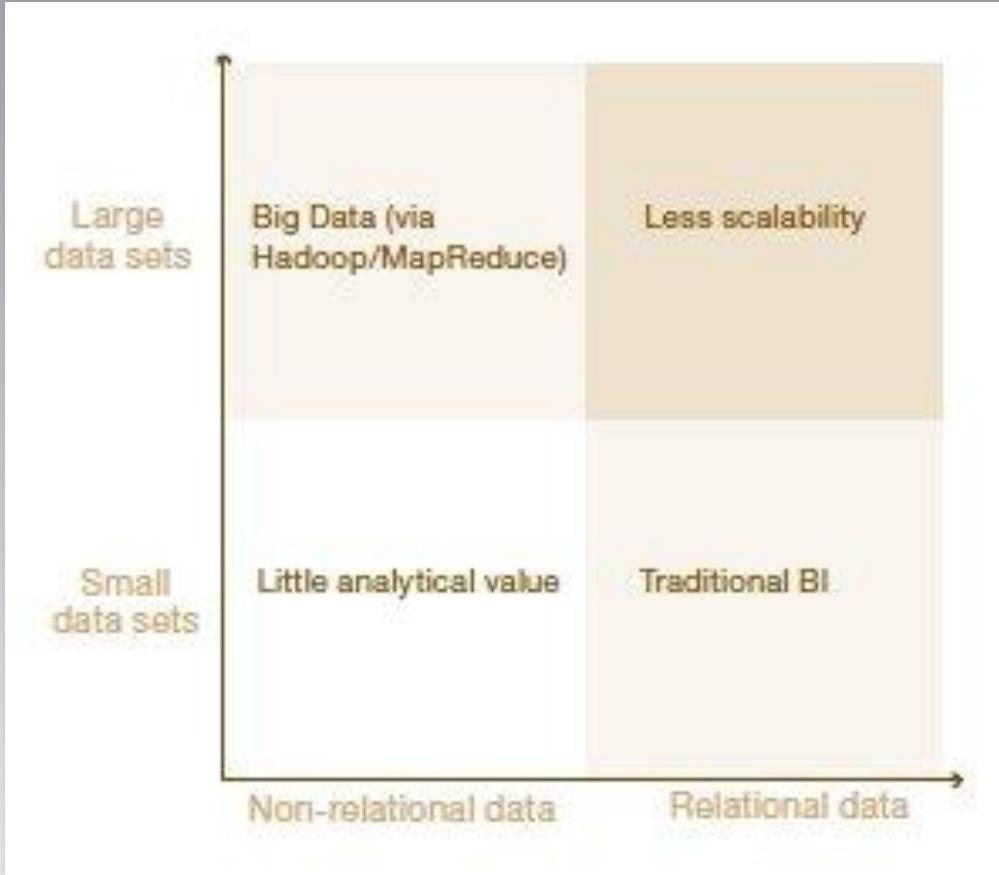
- ◆ 통계, 예측, 최적화를 기반으로 향상된 계획과 의사결정 등을 지원하기 위해 정보를 지식 수준의 형태로 변환하는 프로세스와 도구



Thomas H. Davenport and Jeanne G. Harris



비즈니스 분석과 고급분석 무엇이 다른가?



Galen Gruman,
Executive Editor at InfoWorld
and Principal of the Zango Group

비즈니스 분석과 고급분석 무엇이 다른가?

◆ 비즈니스 분석

- ▶ 현재의 현상과 결과적인 관점을 제시하는데 초점을 맞춤

◆ 고급분석

- ▶ 대용량의 데이터로부터 숨겨진 패턴을 발견하고 상황을 예측

고급분석 기술

- ◆ 비즈니스 상황을 예측하고 효율적인 의사결정을 지원하기 위해 구조화 및 비구조화된 복잡한 형태의 데이터에서 요인들 간의 상관관계와 의미있는 데이터의 패턴을 식별하고 예측하기 위한 모든 기법과 기술들

- James Kobielski, Forrester



고급분석 기술의 구분

Neil Raden, Vice President at Constellation Research



최적화

(Optimization)

예측분석

(Predictive Analytics)

기술분석

(Descriptive Analytics)

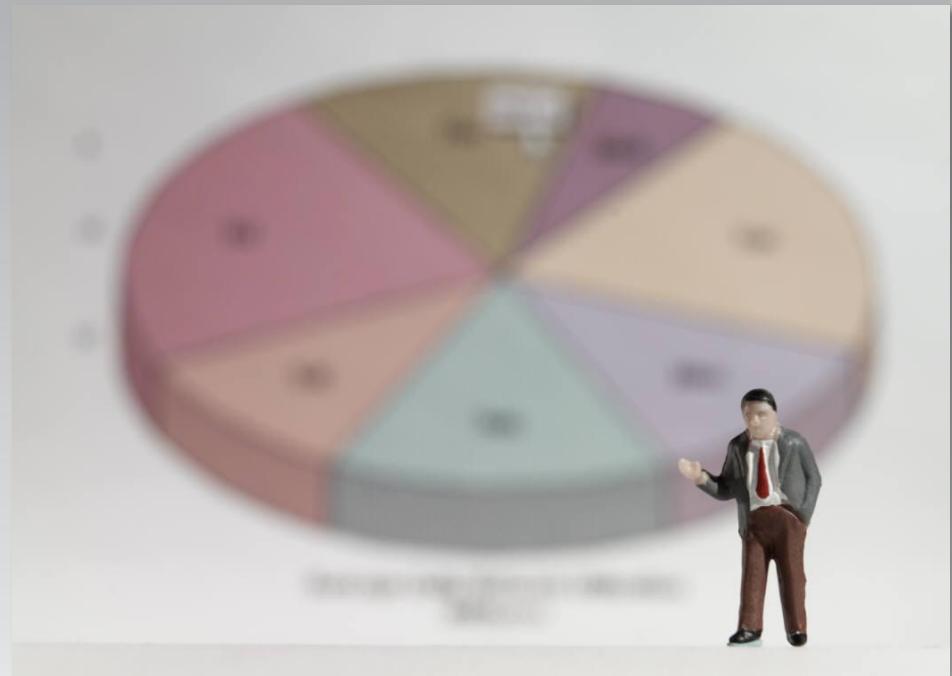
기술분석(Descriptive Analytics)

◀ 기술분석의 목표

- ◆ 과거에서부터 현재까지 주어진 데이터로부터 현재의 상황을 설명할 수 있는 패턴(Pattern)을 찾아 사용자의 이해를 돋기 위해 표현하거나 설명하는 것
 - Jonathan McGrew
- ◆ 즉, 주어진 기간 동안에 어떤 사건이 얼마나 많이 혹은 얼마나 자주 어디에서 발생했는지에 대한 정보를 제공

기술분석을 위한 분석 기법

- ◆ Association Analysis
- ◆ Clustering
- ◆ Classification
- ◆ ...



예측분석(Predictive Analytics)

◆ 예측분석의 목표

- ◆ 과거의 데이터나 사건으로부터 미래에 발생 가능한 상황이나 사건을 예측하여 선제적인 의사결정을 지원

- Charles Nyce, Senior Director of AICPCU



예측분석을 위한 모형과 알고리즘

Analysis	Model Type	Algorithm
Classification	Decision Tree	CART, CHAID, C4.5
	Memory-Based Reasoning	k-nearest neighbor
	Bayesian Classification	Naive Bayes
Clustering	Partition Based	k-means, k-medoids, Expectation Maximization
	Hierarchical Based	BIRCH, CURE
	Navigation Sequencing	Markov Chain
Association	Market Basket Analysis	Association Rules
	Sequence Discovery	Exponential Smoothing, Box-Jenkins
	Link Analysis	Directed Graphs
Estimation	Regression	Linear, Logistic, Supervised Learning
	Neural Networks	Backward Propagation, Feed Forward, Genetic Algorithm
Description	Exploratory Data Analysis Dimensionality Reduction	Histograms, Box-and-Whisker, Pareto Principal Components, Factor Analysis <small>Kent Bauer, Managing Director at GRT Corporation</small>

예측분석을 위한 최근의 경향

- ◆ 보다 쉽게 예측모형을 개발하고 적용할 수 있도록 예측분석을 위한 도구와 솔루션
- ◆ 내부 시스템으로써 운용이 가능한 형태의 모형
- ◆ 많은 형태의 비구조화 데이터를 지원하는 기능
- ◆ 여러 도구들이 오픈 소스(Open Source)로 변모



Dr. Fern Halper, Partner Hurwitz & Associates

예측분석을 위한 기술적 요구사항

- ◆ 그래픽 기반의 모델링이 가능한 분석 작업 도구
- ◆ 둘 이상의 모형이나 알고리즘을 자동적으로 실행함으로써 모형 개발자의 생산성을 높이는 기능
- ◆ 높은 성능의 분석
- ◆ 비정형 문서에 대한 분석 기능

Prof. Marcus Hudec, Universitat Wien



최적화(Optimization)

- ◆ 주어진 가능한 결과들에 대한 평가를 수행하여 최적의 결과를 도출하는 것
- ◆ 최적화 분석을 위한 요구사항
 - ◆ 비즈니스 환경에서 취할 수 있는 여러 가지 대안들 중 제시된 전략을 평가하고 최적의 대안을 선택하도록 도와줄 수 있는 분석 기법이 필요
 - ◆ 확률적 기법 및 통계적 기법과 함께 기술모형과 예측모형을 병합

그 외의 고급분석 이슈

- ◆ 컨텐트 분석(Content Analysis)
- ◆ 텍스트 분석(Text Analytics)
- ◆ 실시간 분석(Realtime Analytics)

컨텐트 분석(Content Analysis)

- ◆ 디지털 환경에서 생성되는 정형 및 비정형을 포함하여 여러 수준의 컨텐트를 비즈니스 인텔리전스와 비즈니스 전략의 가치를 높이기 위한 하나의 방법
- ◆ 컨텐트 분석의 목표
 - ◆ 보다 향상된 의사결정을 지원하기 위한 트랜드나 패턴을 발견하는 것

컨텐트 분석의 활용

Purpose	Element	Question	Use
Make inferences about the antecedents of communications	Source	Who?	<ul style="list-style-type: none"> Answer question of disputed authorship
	Encoding process	Why?	<ul style="list-style-type: none"> Secure political & military intelligence Analyse traits of individuals Infer cultural aspects & change Provide legal & evaluative evidence
Describe & make inferences about the characteristics of communications	Channel	How?	<ul style="list-style-type: none"> Analyse techniques of persuasion Analyse style
	Message	What?	<ul style="list-style-type: none"> Describe trends in communication content Relate known characteristics of sources to messages they produce Compare communication content to standards
	Recipient	To whom?	<ul style="list-style-type: none"> Relate known characteristics of audiences to messages produced for them Describe patterns of communication
Make inferences about the consequences of communications	Decoding process	With what effect?	<ul style="list-style-type: none"> Measure readability Analyse the flow of information Assess responses to communications

Ole Holsti, Duke University

웹에서의 비정형 데이터

- ◆ 최근 웹에서 생성되는 데이터의 80%가 비구조화된 데이터
 - Charles Nyce, Senior Director of AICPCU



20%



80%

텍스트 분석(Text Analytics)

- ◆ 비구조화된 데이터로부터 의미 있는 정보를 추출하기 위한 언어적 혹은 통계적 기술
- ◆ 자연어 처리나 여러 분석 방법론을 통해 분석에 활용될 수 있는 형태의 데이터로 변환하는 역할을 포함



<http://spotfireblog.tibco.com/?cat=179>



텍스트 분석의 하위 컴포넌트

- ◆ Information Retrieval or Identification of a Corpus
- ◆ Natural Language Processing
- ◆ Named Entity Recognition
- ◆ Recognition of Pattern Identified Entities
- ◆ Coreference
- ◆ Relationship, Fact, and Event Extraction
- ◆ Sentiment Analysis
- ◆ Quantitative Text Analysis

텍스트 분석의 적용

- ◆ 전사적 비즈니스 인텔리전스/데이터 마이닝
- ◆ 국가 보안
- ◆ 생명 과학
- ◆ 감성 분석도구 및 플랫폼
- ◆ 자연어 처리 도구와 서비스
- ◆ 자동화된 광고 노출
- ◆ 정보검색 및 처리
- ◆ 소셜 미디어 분석



[http://www.greenbookblog.org/2012/01/02/
from-sentiment-analysis-to-enterprise-applications/](http://www.greenbookblog.org/2012/01/02/from-sentiment-analysis-to-enterprise-applications/)

실시간 분석(Realtime Analytics)

- ◆ 분석에 필요한 모든 가능한 데이터를 활용하여 사용자가 분석을 수행하고 하는 시점에 빠르고 적시에 지식을 제공해 줄 수 있는 분석 기법
- ◆ 정확성보다는 얼마나 빠르게 사용자가 원하는 시점에 적절히 제공해 줄 수 있는지에 초점이 맞추어져 있음



<http://www.optify.net/social-media/how-realtime-marketers-can-leverage-promoted-tweets/>

실시간 분석을 위한 지원기술

- ◆ 데이터베이스 자체에 분석 로직을 포함하여 분석을 수행하기 위한 인-데이터베이스 분석
- ◆ 분석 처리를 위해 만들어진 하드웨어와 소프트웨어를 결합하여 활용하는 데이터 웨어하우스
- ◆ 빠른 데이터 처리를 위해 메모리의 인덱스를 활용하는 인-메모리 분석
- ◆ 다중 프로세스를 활용하는 MPP(Massively Parallel Programming)

목 차

I. 빅 데이터

II. 고급분석 기술 동향

III. 고급분석 지원을 위한 인프라

IV. 향후 과제

신속한 분석을 위한 기술

◆ 인-데이터베이스 기술

- 데이터베이스와 분석 소프트웨어의 분리로 인한 데이터의 처리 및 프로세스 등의 여러 단계를 거치지 않고 보다 분석 시점에 신속하게 데이터를 분석



신속한 분석을 위한 기술

◆ 인-메모리 기술

- ❖ 디스크 대신 메모리를 이용하여 색인을 만들고 데이터를 처리
- ❖ 데이터 모형을 만들고 질의를 분석하며 다양한 관점의 분석을 처리하는데 소요되는 시간을 줄임
- ❖ 그러나 하드웨어에 대한 많은 투자를 필요로 함

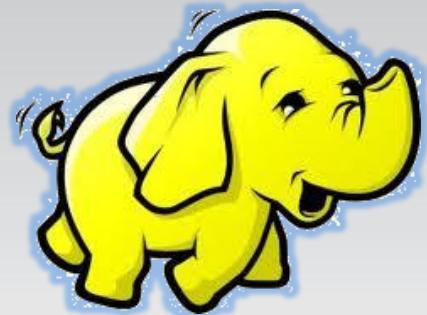


Oracle's Exalytics In-Memory Machine

분산처리를 위한 기술

◆ 아파치 하둡(Hadoop)

- ❖ 대규모의 컴퓨터 클러스터에서 동작하는 분산 애플리케이션 개발을 위한 자바 오픈소스 프레임워크
- ❖ 하둡 분산 파일 시스템
 - ❖ 하둡 프레임워크를 기반으로 자바로 구현된 분산 파일 시스템
- ❖ 맵리듀스
 - ❖ 클러스터를 구성하고 대용량의 데이터를 클러스터를 구성하는 각각의 컴퓨터에 분산시켜 처리할 수 있는 프레임워크



분산처리를 위한 기술

◆ 분산 데이터베이스

- ❖ 물리적으로 분리된 두 대 이상의 컴퓨터에서 데이터를 분산시켜 저장하고 처리할 수 있는 데이터베이스
- ❖ NoSQL
 - ❖ 대용량의 비정형 데이터를 테이블 구조가 아닌 다른 형태로 분산 저장하여 처리



<http://www.pentaho.com/big-data/nosql/>

분산처리를 위한 기술

◆ 분산 데이터베이스

- ❖ NoSQL 데이터베이스의 형태
 - ❖ Key/Value 데이터베이스
 - ❖ 빅테이블
 - ❖ 문서 데이터베이스
 - ❖ 그래프 데이터베이스



<http://www.dataversity.net/nosql-job-of-the-day-senior-java-programmer/4026/>

목 차

- I . 빅 데이터
- II . 고급분석 기술 동향
- III . 고급분석 지원을 위한 인프라
- IV . 향후 과제

향후 과제

◆ 빅 데이터 분석의 적용 분야

- ◆ 경쟁환경에서의 우위를 선점할 수 있는 선제적 의사결정
- ◆ 사건에 대한 징후와 경과를 파악
- ◆ 객관적인 분석을 통해 효과적인 의사결정을 촉진
- ◆ 전략을 실행함으로써 발생하는 효과를 예측

향후 과제

◆ 빅 데이터 분석을 위한 해결과제

- ◆ 전문적으로 분석 모형을 개발하고 수행할 수 있는 전문인력의 필요
- ◆ 분석을 수행하는데 필요한 투자에 적극적으로 동참
- ◆ 개인의 프라이버시 문제

Q & A

Dr. Myungjin Lee

e-Mail : xml@yonsei.ac.kr

Twitter : <http://twitter.com/MyungjinLee>

Facebook : <http://www.facebook.com/mjinlee>

SlideShare : <http://www.slideshare.net/onlyjiny/>

감사합니다.

