



고려대학교

빅데이터 기반 지능정보시스템 개발 과정

빅데이터 기반 지능정보 시스템






■ 데이터의 규모에 초점을 맞춘 정의

- 일반적인 데이터베이스 소프트웨어가 데이터 수집, 저장, 관리, 분석할 수 있는 역량을 넘어서는 규모의 데이터 (McKinsey, 2011)

■ 업무수행 방식에 초점을 맞춘 정의

- 다양한 종류의 대규모 데이터로부터 저렴한 비용으로 가치를 추출하고, 데이터의 빠른 수집, 발굴, 분석을 지원하도록 고안된 차세대 기술 및 아키텍처 (IDC-International Data Corporation, 2011)

Data sizes

	Examples	Characteristics	Typical tools	Analytical methods
 Small Data (megabytes)	Sales records, Customers database (small and medium companies)	Hundreds – thousands of records	Personal computer, Excel, R, other basic statistics software	Simple statistics
 Large Data (gigabytes-terabytes)	Customer databases (big companies)	Millions of records, mostly <u>structured</u> data	Server workstation computer, <u>Relational</u> database systems, data warehouses	Advanced statistics, business intelligence, data mining,
 Big Data (terabytes – petabytes)	Customer interactions (social media, mobile), multimedia (video, images, free text), location-based data, RFIM	Over millions of records, <u>distributed</u> , <u>unstructured</u>	Cloud, data centers, <u>Distributed</u> databases, NoSQL, Hadoop	MapReduce, Distributed File Systems

■ 빅데이터의 가치

○ 경제적 가치의 원천

- ▶ 빅데이터에서 유용한 정보를 찾아내고, 잠재된 정보를 활용할 수 있는 기업들이 경쟁에서 시장을 선도

○ 의사결정 속도 개선

- ▶ 데이터 지향적 마케팅에 대한 전략적 접근은 기업의 혁신, 경쟁력, 생산성 등 모든 영역에 긍정적인 영향

○ 새로운 가치 창출

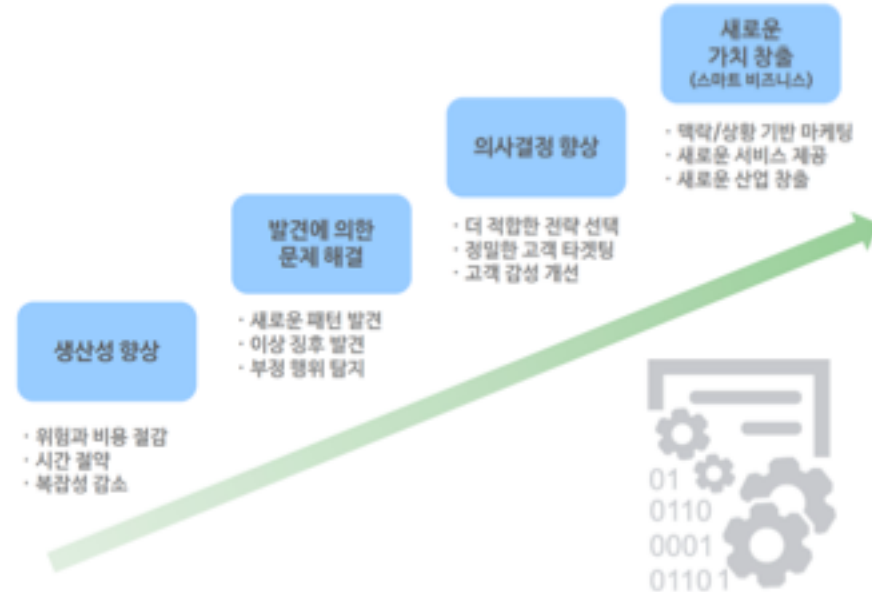
- ▶ 데이터 기반의 의사결정은 조직의 새로운 기회, 채널 또는 시장의 창출, 비즈니스 수행 방식에 변화

데이터 분석 활용 가치

기관명	주요 내용
Economist(2010)	<ul style="list-style-type: none"> · 데이터는 자본이나 노동력과 거의 동등한 레벨의 경제적 투입 자본, 비즈니스의 새로운 원자재 역할 · 비즈니스 트렌드 파악, 질병 예방, 범죄 해결 등 효과
MIT Sloan(2010)	<ul style="list-style-type: none"> · 데이터 분석을 잘 활용하는 조직일수록 차별적 경쟁력을 갖추고 높은 성과를 창출 · 조직 분석역량 3단계(열망-숙련-변혁 단계) 특징 제시
PwC(2010)	<ul style="list-style-type: none"> · 빅데이터는 이전까지는 다루지 못하고, 시도하지 못했던 데이터의 활용을 가능하게 하여 잠재적 가치와 영향력이 높음 · 빅데이터의 중요성에 대해 기업들이 주목하고 있으며, 새로운 비즈니스의 가치 창출의 핵심 키가 될 것
Gartner(2011)	<ul style="list-style-type: none"> · 데이터는 21세기 원유, 데이터가 미래 경쟁 우위를 좌우 · 기업은 다가올 '데이터 경제 시대'를 이해하고 정보 고립(Information Silo)을 경계해야 성공 가능 · 빅데이터는 향후 주목해야 할 이머징 기술(2-5년 후 성숙)
Mckinsey(2011)	<ul style="list-style-type: none"> · 글로벌 비즈니스 지형을 뒤바꿀 기술 트렌드의 3가지 핵심은 '클라우드', '빅데이터', '스마트 자산(Smart assets)' · 빅데이터는 혁신, 경쟁력, 생산성의 핵심 요소 · 의료, 공공행정 등 5대 분야에서 6천억 달러 이상 가치 창출

출처: Economist, MIT Sloan, PwC, Gartner, McKinsey, NIA(2011.12), **새 가치창출 엔진**, 빅데이터의 새로운 가능성과 대응전략 재인용

데이터 분석에 의한 혁신 단계

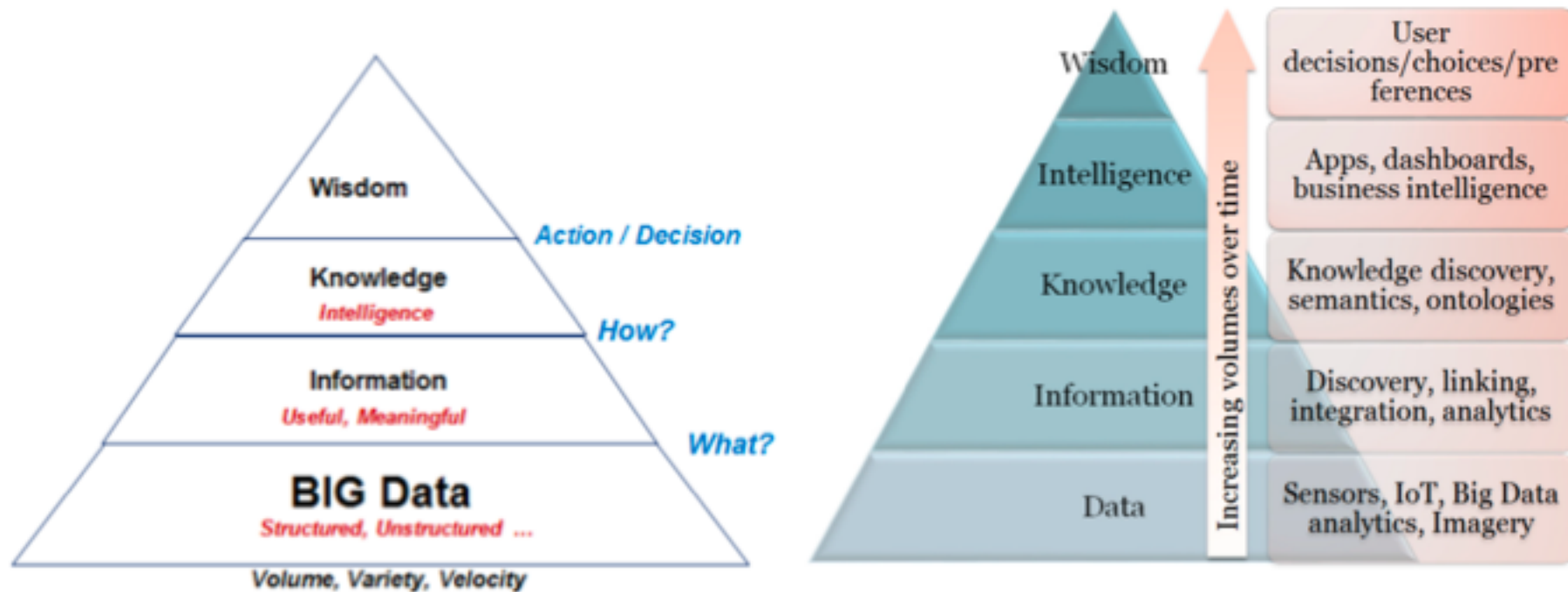





출처: 함유근, 채승범(2012), 빅데이터, 경영을 바꾸다

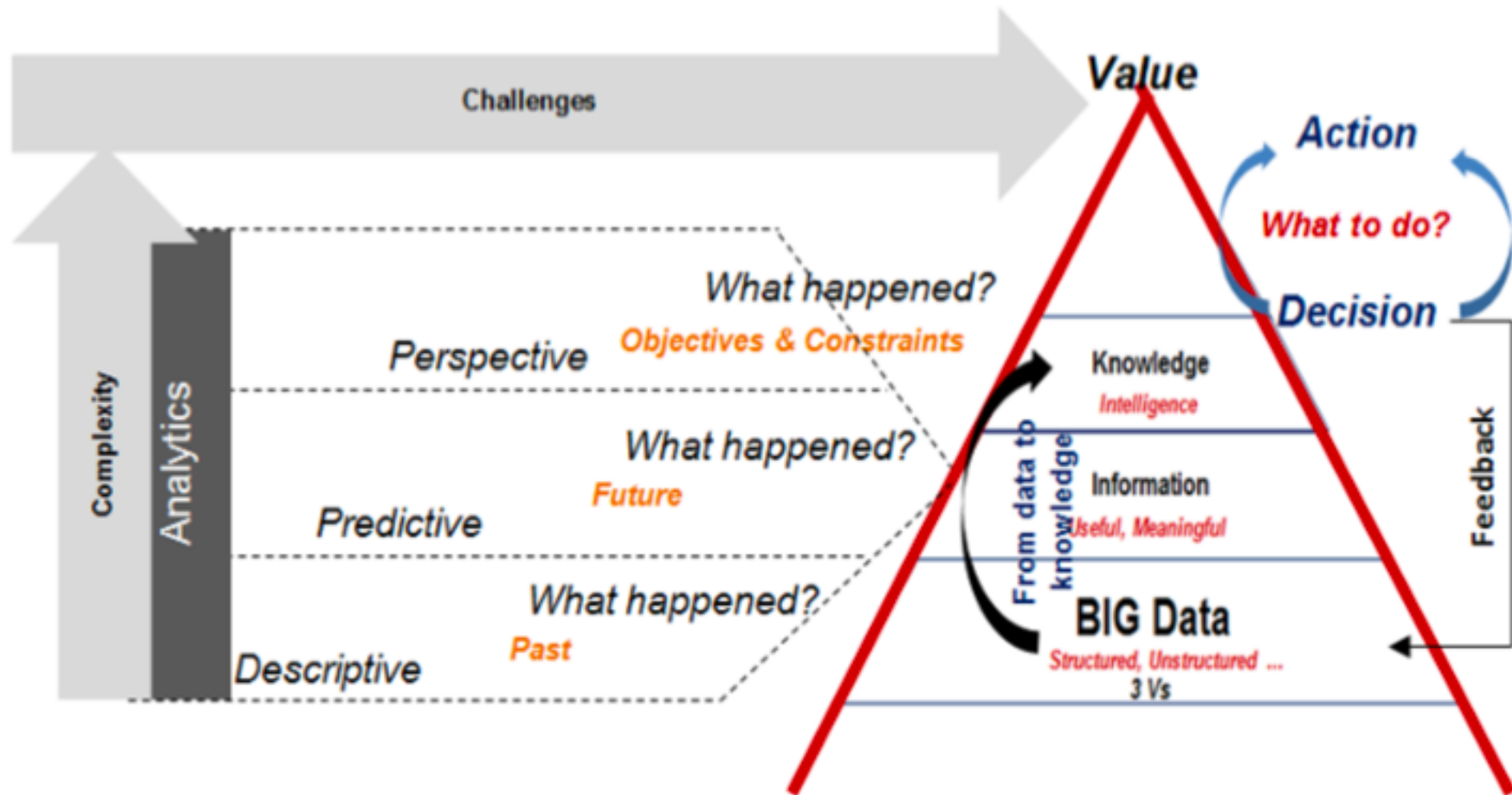
■ 빅데이터의 역할

미래사회 특징		빅데이터의 역할
불확실성	통찰력	<ul style="list-style-type: none"> - 사회현상, 현실세계의 데이터를 기반으로 한 패턴분석과 미래 전망 - 여러 가지 가능성에 대한 시나리오 시뮬레이션 - 다각적인 상황이 고려된 통찰력을 제시 - 다수의 시나리오의 상황 변화에 유연하게 대처
리스크	대응력	<ul style="list-style-type: none"> - 환경, 소셜, 모니터링 정보의 패턴 분석을 통한 위험징후, 이상 신호 포착 - 이슈를 사전에 인지, 분석하고 빠른 의사결정과 실시간 대응지원 - 기업과 국가 경영의 명성 제고 및 낭비요소 절감
스마트	경쟁력	<ul style="list-style-type: none"> - 대규모 데이터 분석을 통한 상황 인지, 인공지능 서비스 가능 - 개인화, 지능화 서비스 제공 확대 - 소셜분석, 평가, 신용, 평판 분석을 통해 최적의 선택 지원 - 트렌드 변화 분석을 통한 제품 경쟁력 확보
융합	창조력	<ul style="list-style-type: none"> - 타 분야와의 융합을 통한 새로운 가치 창출 - 인과관계, 상관관계가 컨버전스 분야의 데이터 분석으로 안정성 확보, 시행착오 최소화 - 방대한 데이터 활용을 통한 새로운 융합시장 창출

빅데이터 기반 **지능정보** 시스템



		 Tools used	 Limitations	 When to use
Descriptive Analytics	<i>What happened and why?</i>	<ul style="list-style-type: none"> › Data aggregation › Data mining 	<ul style="list-style-type: none"> › Snapshot of the past › Limited ability to guide decisions 	<ul style="list-style-type: none"> › When you want to summarize results for all/part of your business
Predictive Analytics	<i>What might happen?</i>	<ul style="list-style-type: none"> › Statistical models › Simulation 	<ul style="list-style-type: none"> › Guess at the future › Helps inform low complexity decisions 	<ul style="list-style-type: none"> › When you want to make an educated guess at likely results
Prescriptive Analytics	<i>What should we do?</i>	<ul style="list-style-type: none"> › Optimization models › Heuristics 	<ul style="list-style-type: none"> › Most effective where you have more control over what is being modeled 	<ul style="list-style-type: none"> • When you have important, complex or time-sensitive decisions to make



빅데이터 기반 **지능정보 시스템**

■ 아마존

- 소비자가 읽었던 도서 목록 자료를 분석해 새로운 책들을 제작, 추천하는 방식을 시행
- 회원들의 소비 패턴을 분석해 구매 가능한 상품을 추천, 아마존 매출의 35%가 추천 상품에서 발생

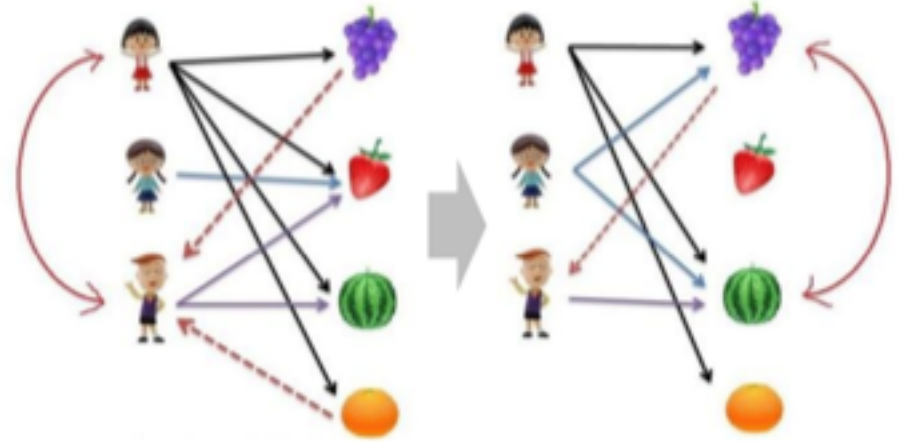
아마존 웹사이트



▲ 아마존 메인 화면, 아마존은 빅데이터를 이용한 분석 기술을 토대로 고객에게 맞춤형 추천 상품을 제공함.

출처: Amazon

아마존의 Item-to-item collaborative filtering



[유저 기반 필터링] **[아이템 기반 필터링]**

▲ 기존의 데이터를 통해 상품을 추천해주는 알고리즘으로 고객 수나 아이템 수와 관계없이 대량의 데이터를 처리해서 고품질의 추천을 해줌. 상품 간의 상관관계를 결정하는 아이템 매트릭스를 만든 후 고객의 최신 입력 데이터를 기반으로 고객의 기호를 유추해서 상품을 추천.

즉, 기존의 온라인 쇼핑몰에서 자주 활용하던 소비자의 패턴 분석에 의한 추천 방식이 아닌 구매한 물건 혹은 서치한 물건 중심으로 추천하는 방식.

출처: <http://blog.naver.com/sanny0314/220630765408>

■ 넷플릭스

○ 사용자의 성향 파악 알고리즘 개발

- 사용자 별점, 위치정보, 기기정보, 플레이 수, 평일/주말 선호 프로그램, 소셜 미디어 내에서 언급된 횟수 등을 분석
- 이 밖에도 시청률 조사업체, 기타 시장조사업체들이 제공하는 메타데이터, 소셜 미디어에서 수집한 소셜 데이터에 이르기까지 모두 수집 분석 시청자의 성향을 파악함.

○ 하우스 오브 카드(House of Cards) 제작 방영

- 이용자의 선호도 분석을 통해 원하는 드라마, 원하는 배우와 감독, 원하는 스토리 추천

넷플릭스 웹사이트

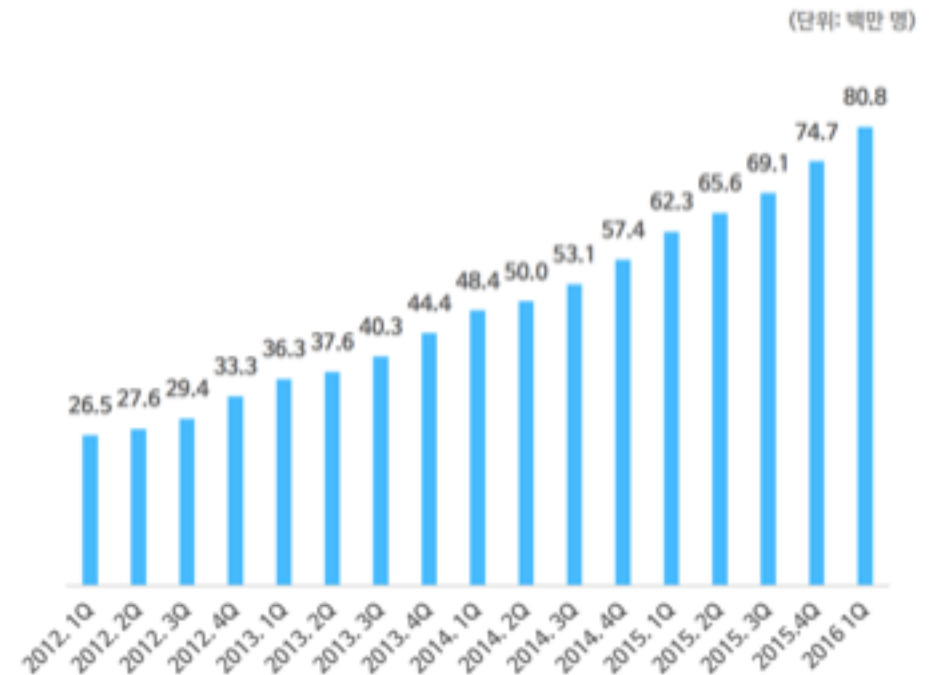


▲ '하우스 오브 카드(House of Cards)'로 큰 성공을 거둔 넷플릭스 웹사이트

보통 방송사, 영화사들은 신규 프로그램을 방영할 경우 막대한 광고를 송출. 반면, 3천 700만 명이 넘는 회원을 확보하고 있는 넷플릭스에서는 광고 대신 빅데이터를 활용. 빅데이터 분석에 기반한 넷플릭스 콘텐츠 추천 서비스는 드라마 제작에 큰 영향을 미침.

출처: 정보통신정책연구원(2014.12), 넷플릭스의 빅데이터

넷플릭스의 분기별 가입자 증가 추이



출처: 2016년은 예상

출처: Netflix



고려대학교

빅데이터 기반 지능정보 시스템 개발 과정

7주차, 8주차, 9주차 : 빅데이터 기반의 지능정보시스템 개발

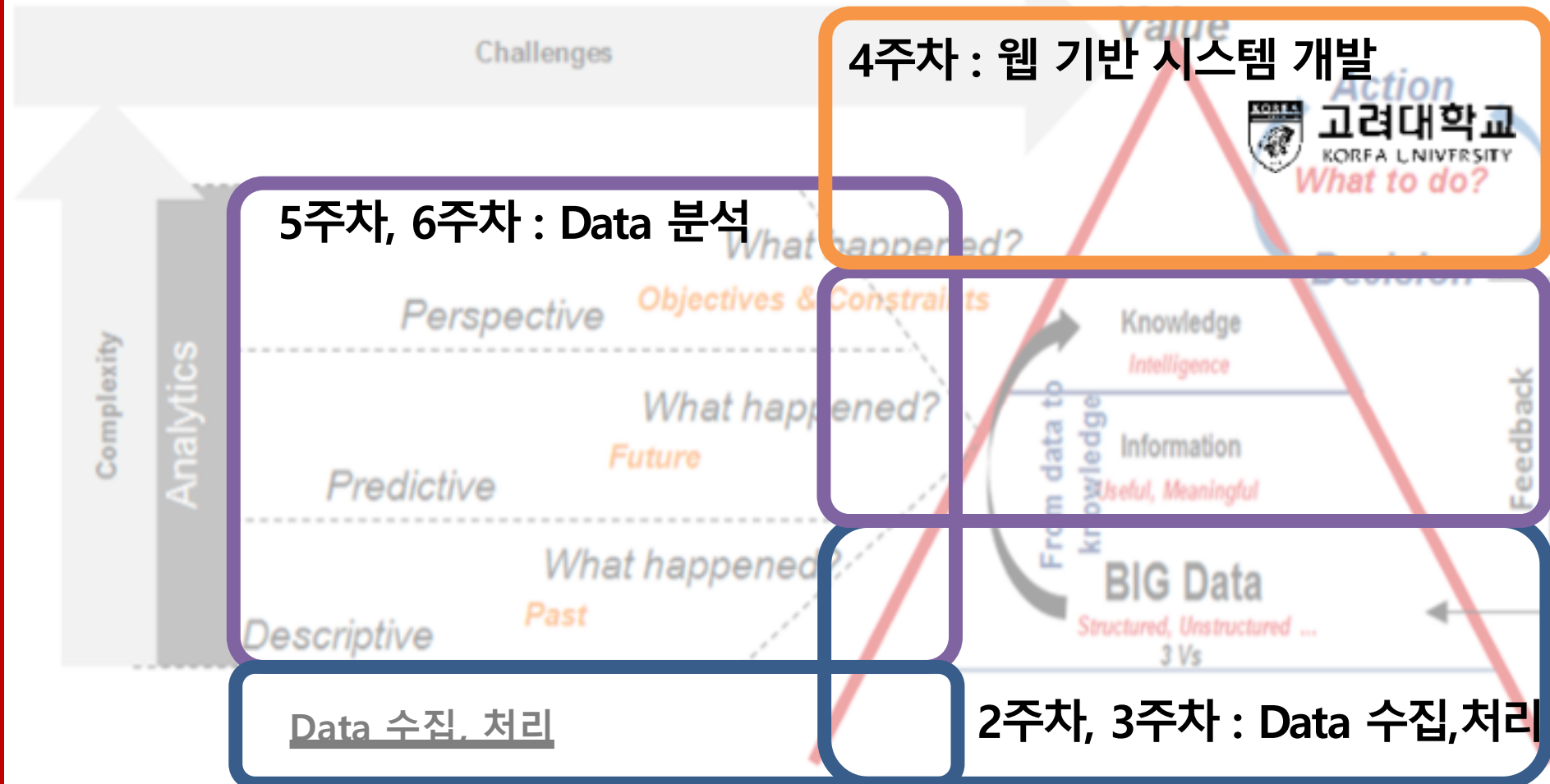
4주차 : 웹 기반 시스템 개발



5주차, 6주차 : Data 분석

Data 수집, 처리

2주차, 3주차 : Data 수집, 처리

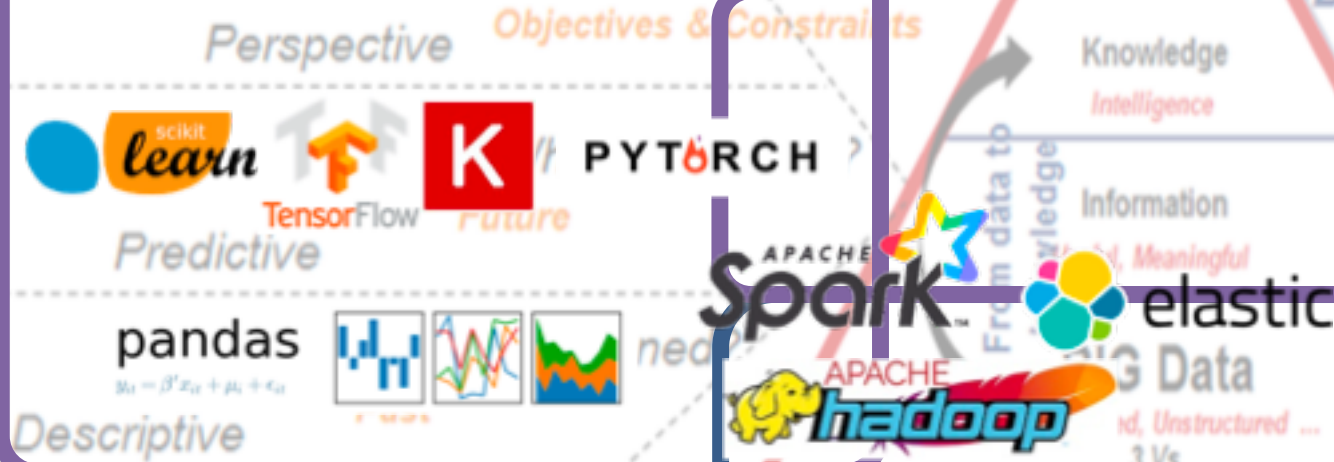


7주차, 8주차, 9주차 : 빅데이터 기반의 지능정보시스템 개발

4주차 : 웹 기반 시스템 개발



5주차, 6주차 : Data 분석

Data 수집, 처리

2주차, 3주차 : Data 수집,처리

GUI vs Java vs R vs Python

Non (little)
Programming

Programming

분석 등 기존 기법에 대해서 접근 용이성
but 비용 증가, 자신의 기법 등 추가의 어려움

Java vs R vs Python

compile

interpreter

시스템의 안정화
but 개발 생산성이 좋지 않음



VS



Domain Specific

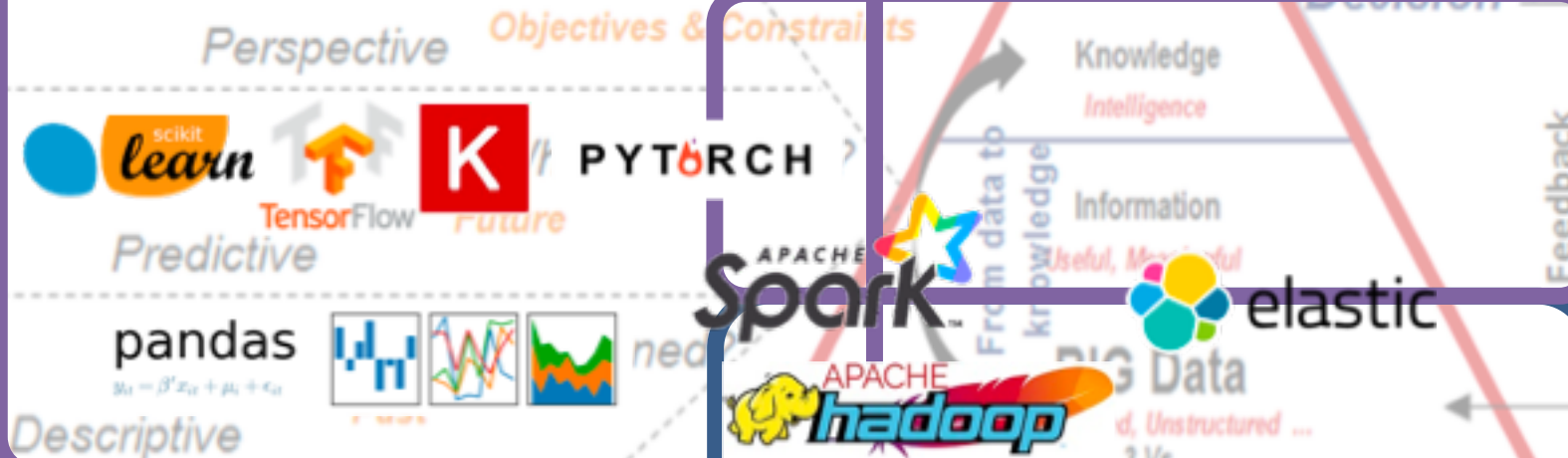
**General Purpose
Porting의 용이성**

7주차, 8주차, 9주차 : 빅데이터 기반의 지능정보시스템 개발

4주차 : 웹 기반 시스템 개발



5주차, 6주차 : Data 분석

Data 수집, 처리

2주차, 3주차 : Data 수집, 처리

1주차 : Python



■ 수업/실습

- 빅데이터 기반 지능정보 시스템을 개발하기 위한 **필수적인** 지식 습득 (개념, 도구 사용 등) 위주
- 실습 위주로 수업 진행 / 매일 1-2시간씩 과제 풀이(복습 문제 위주), Help 세션 진행

■ 프로젝트

- 개인별 관심있는 분야에 대해서 특화해서 진행
 - Git을 활용한 포트폴리오 제작
 - 강사, 조교, 산업계 등 멘토링
- 개인 프로젝트
 - 1주차, 2주차, 3주차 : Python 라이브러리를 활용한 빅데이터 수집, 처리 등
- 미니 팀 프로젝트 (역할 분담)
 - 4주차 : Flask 기반의 웹 시스템 개발
- 미니 팀 프로젝트 (도메인)
 - 5주차, 6주차 : 빅데이터 분석
 - pandas, spark mlib, scikit-learn, tensorflow(keras,pytorch) / matplotlib, geopandas, folium 등의 라이브러리를 활용
 - 관심있는 도메인(데이터진흥원 제공 데이터, 공공데이터 포털내 데이터, Kaggle 데이터)에 대해서 팀을 구성후 데이터 분석
- 팀 프로젝트 (역할 분담 + 도메인)
 - 7주차,8주차,9주차 : 빅데이터 기반의 지능정보 시스템 개발