Data Science

데이터 과학 개요

강사 양석환

데이터 개요

AiDA Lab.

데이터란 무엇인가?

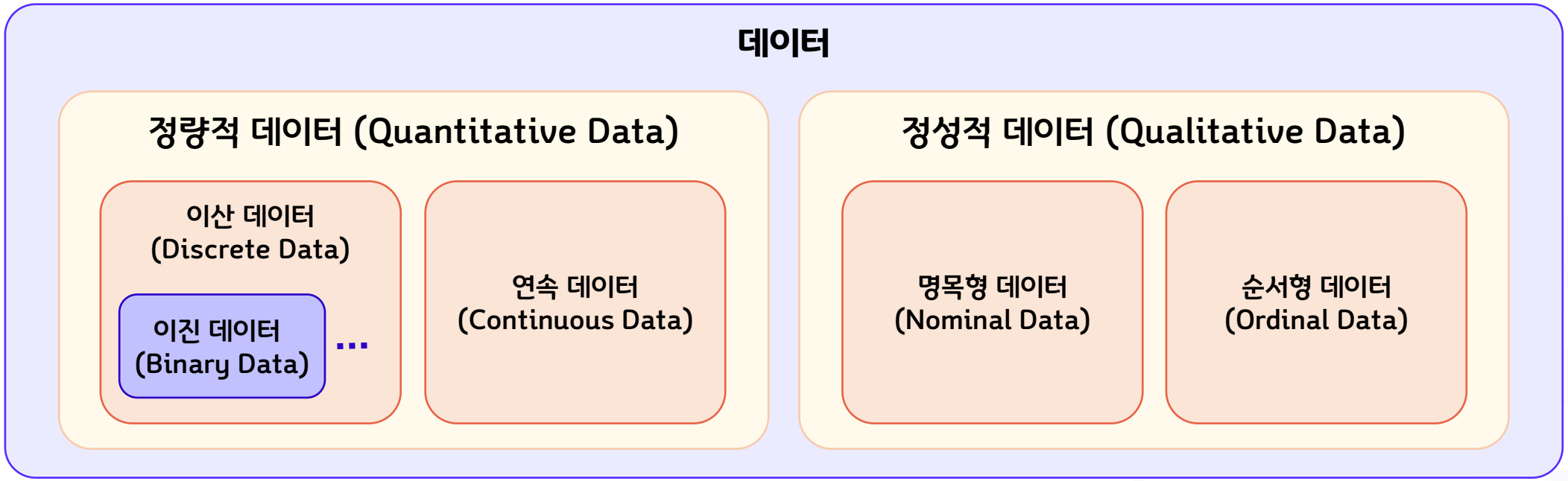
- **데이터의 정의**
  - **추론과 추정의 근거를 이루는 사실 (옥스퍼드 대사전)**
  - **데이터는**
    - **객관적 사실로서 추론, 예측, 전망, 추정을 위한 근거로 작용하는 것**
    - **각각의 개별 데이터는 그 자체로 큰 의미가 없음 → 정보로 변환 필요**
  - **정보란 데이터가 특정 기준에 따라 가공, 처리 및 분류되고 정리되어 데이터 간 연관관계 속에서 의미를 가지며, 유용한 효과를 가지도록 한 것**
- **데이터의 분류**
  - **기준에 따라 매우 다양한 분류가 존재함**

이후의 데이터 분석, 데이터 마이닝,
인공지능 모델 등을 다루기 위해서는
데이터의 정의 및 분류 등에 대한 개념과
관련 용어들을 잘 이해해 두는 것이 좋음

Artificial Intelligence & Data Analysis Laboratory

데이터란 무엇인가?

- **데이터의 속성에 따른 분류**



정량적 데이터/양적자료 (Quantitative Data)
= 수치형 데이터 (Numerical Data)

정성적 데이터/질적자료 (Qualitative Data)
= 범주형 데이터 (Categorical Data)

Artificial Intelligence & Data Analysis Laboratory

데이터란 무엇인가?

- **데이터의 유형에 따른 분류**
  - **정형 데이터 (Structured Data)**
    - 정해진 형식과 구조에 맞게 저장되도록 구성된 데이터. 연산 가능
    - 예: 관계형 데이터베이스의 테이블에 저장되는 데이터 등
  - **반정형 데이터 (Semi-structured Data)**
    - 데이터의 형식과 구조가 비교적 유연하고 스키마 정보를 데이터와 함께 제공하는 파일 형식의 데이터. 연산 불가능
    - 예: JSON, XML, RDF, HTML 등
  - **비정형 데이터 (Unstructured Data)**
    - 구조가 정해지지 않은 대부분의 데이터. 연산 불가능
    - 예: 동영상, 이미지, 음성, 문서, 메일 등

AiDA Lab.
Artificial Intelligence & Data Analysis Laboratory

데이터란 무엇인가?

- **정량적 데이터와 정성적 데이터의 종합적 비교**

	정량적 데이터	정성적 데이터
유형	정형 데이터, 반정형 데이터	비정형 데이터
특징	여러 요소의 결합으로 의미 부여	객체 하나가 함축된 의미를 내포함
관점	주로 객관적인 내용	주로 주관적인 내용
구성	수치나 기호 등	문자나 언어 등
형태	데이터베이스, 스프레드시트 등	웹 로그, 텍스트 파일 등
위치	DBMS, 로컬 시스템 등 시스템의 내부에 위치	웹사이트, 모바일 플랫폼 등 시스템의 외부에 위치
분석	통계 분석 시 용이함	통계 분석 시 어려움

Artificial Intelligence & Data Analysis Laboratory

데이터란 무엇인가?

- **데이터의 근원에 따른 분류**
  - **데이터 수집과정은 데이터의 재생산 과정**
  - **가역 데이터**
    - 생산된 데이터의 원본으로 일정 수준 환원이 가능한 데이터
    - 원본과 1:1 대응 관계 → 환원 가능 → 이력추적 가능 → 원본 데이터가 변경되는 경우 변경사항 반영 가능
  - **불가역 데이터**
    - 생산된 데이터의 원본으로 환원이 불가능한 데이터
    - 재생산 시, 원본 데이터와는 전혀 다른 형태로 재생산됨 → 환원 불가

AiDA Lab.
Artificial Intelligence & Data Analysis Laboratory

데이터란 무엇인가?

- **가역 데이터와 불가역 데이터의 종합적 비교**

	가역 데이터	불가역 데이터
환원성(추적성)	가능 (비가공 데이터)	불가능 (가공 데이터)
의존성	원본 데이터 그 자체	원본 데이터와 독립된 새 객체
원본과의 관계	1:1의 관계	1:N, N:1, 또는 M:N의 관계
처리 과정	탐색	결합
활용 분야	데이터 마트, 데이터 웨어하우스 등	데이터 전처리, 프로파일 구성 등

Artificial Intelligence & Data Analysis Laboratory

빅데이터 개요

AiDA Lab.

빅데이터란?

- 빅데이터의 정의
  - 일반적인 데이터베이스 소프트웨어로 저장, 관리, 분석할 수 있는 범위를 초과하는 규모의 데이터 (McKinsey, 2011)
  - 다양한 종류의 대규모 데이터로부터 저렴한 비용으로 가치를 추출하고 데이터의 초고속 수집, 발굴, 분석을 지원하도록 고안된 차세대 기술 및 아키텍처 (IDC, 2011)
  - 대용량 데이터를 활용해 작은 용량에서는 얻을 수 없었던 새로운 통찰이나 가치를 추출해 내며, 나아가 이를 활용해 시장과 기업 및 시민과 정부의 관계 등 많은 분야에서 변화를 가져오는 것 (Mayer-Schonberger & Cukier, 2013)

Artificial Intelligence & Data Analysis Laboratory

빅데이터의 등장 배경

- PC와 인터넷, 모바일 기기의 이용이 생활화되면서 빅데이터 환경이 조성되기 시작함
  - 데이터의 관점에서 보면 과거에는 상점에서 물건을 살 때만 데이터가 기록되었지만 인터넷 쇼핑몰의 경우에는 구매를 하지 않더라도 방문자가 돌아다닌 기록이 자동적으로 데이터로 저장됨
  - 어떤 상품에 관심이 있는지, 얼마 동안 쇼핑몰에 머물렀는지 등의 다양한 정보가 저장됨
  - 사람들은 쇼핑뿐 아니라 은행, 증권과 같은 금융거래, 교육과 학습, 여가활동, 자료검색과 이메일 등 하루 대부분의 시간을 PC와 인터넷에 할애하고 있으며, 사람과 기계, 기계와 기계가 서로 정보를 주고받는 사물인터넷(IoT)의 확산도 디지털 정보가 폭발적으로 증가하게 되는 이유

Artificial Intelligence & Data Analysis Laboratory

빅데이터의 등장 배경

- 데이터의 증가 속도 뿐만 아니라, 형태와 질에서도 기존과 다른 양상을 보이기 시작
  - 사용자가 직접 제작하는 UCC를 비롯한 동영상 콘텐츠, 휴대전화와 SNS (Social Network Service)에서 생성되는 문자 등
  - 블로그나 SNS에서 유통되는 텍스트 정보는 내용을 통해 글을 쓴 사람의 성향뿐 아니라, 소통하는 상대방의 연결 관계까지도 분석 가능
  - 기타 사진, 동영상, 방송 콘텐츠, CCTV 영상(도로, 공공건물, 아파트 엘리베이터 등) 등 일상생활의 행동 하나하나가 빠짐없이 데이터로 저장되고 있음

Artificial Intelligence & Data Analysis Laboratory

빅데이터의 등장 배경

- **수많은 의도된 데이터 양산**
  - **민간 분야뿐 아니라 공공 분야도 데이터를 양산 중**
  - **센서스(Census)를 비롯한 다양한 사회 조사, 국세자료, 의료보험, 연금 등의 분야에서 데이터가 생산 중**
  - **코로나 19의 확산으로 인한 비대면 환경의 보편화도 데이터 증가를 가속화**

빅데이터 환경에서는 과거에 비해 데이터의 양이 폭증함과 동시에 데이터의 종류도 다양해 졌고,
이러한 환경 변화에 따라 빅데이터를 이용한다면
사람들의 행동은 물론 위치정보와 SNS를 통해 생각과 의견까지 분석하고 예측할 수 있게 됨

빅데이터의 등장으로 인한 변화

- 기업과 학계의 변화
  - 기업
    - 온·오프라인 고객 데이터가 축적되면서 데이터에 숨어있는 가치를 발굴하여 새로운 성장동력으로 활용
  - 학계
    - 인간 게놈 프로젝트, 기후 관찰 등 거대 데이터를 다루는 학문 분야가 확산되면서 필요한 기술 아키텍처 및 분석 기법들이 발전

AIDA Lab.

빅데이터의 등장으로 인한 변화

- **데이터 처리, 분석 분야에서의 변화**
  - **데이터 처리 시점이 사전 처리(pre-processing)에서 사후 처리(post-processing)로 이동**
    - 필요한 정보만 수집하는 기존의 시스템에서 가능한 한 많은 데이터를 모으고 조합하여 정보를 얻는 방식으로 변화
  - **데이터 처리 범주가 표본조사에서 전수조사로 확대**
    - 기술 발전으로 인해 데이터 처리비용 감소 → 표본조사가 아닌 전수조사를 통해 패턴, 정보 발현 방식으로 변화
  - **데이터의 가치 판단 기준이 질(quality)보다 양(quantity)으로 그 중요도가 달라짐**
    - 데이터 양의 증가가 전체적으로 좋은 결과를 산출하기 위한 긍정적인 영향을 미친다는 추론을 바탕으로 그 중요도가 변화
  - **데이터를 분석하는 방향이 이론적 인과관계 중심에서 단순한 상관관계로 변화**
    - 데이터 기반의 상관관계 분석으로 특정 현상의 발생 가능성을 포착하여 대응하는 방식으로 변화

빅데이터의 특징

대분류	소분류	특징	내용
5V	3V	규모(Volume)	- 데이터 양이 급격하게 증가(대용량화) - 기존 데이터 관리 시스템의 성능적 한계 도달
		유형(Variety)	- 데이터의 종류와 근원 확대(다양화) - 정형 데이터 외 반정형 및 비정형 데이터로 확장
		속도(Velocity)	- 데이터 수집과 처리 속도의 변화(고속화) - 대용량 데이터의 신속하고 즉각적인 분석 요구
	+2V	품질(Veracity)	- 데이터의 신뢰성, 정확성, 타당성 보장이 필수 - 고품질의 데이터에서 고수준 인사이트 도출 가능
		가치(Value)	- 대용량의 데이터 안에 숨겨진 가치 발굴이 중요 - 다른 데이터들과 연계 시 가치가 배로 증대

Artificial Intelligence & Data Analysis Laboratory

전통적 데이터와 빅데이터 비교

	전통적 데이터	빅데이터
규모	기가바이트(GB) 이하	테라바이트(TB) 이상
처리단위	시간 또는 일 단위 처리	실시간 처리
유형	정형 데이터	정형+반정형, 비정형 데이터
처리방식	중앙집중식 처리	분산 처리
시스템	Relational DBMS	Hadoop, HDFS, Hbase, NoSQL 등

Artificial Intelligence & Data Analysis Laboratory

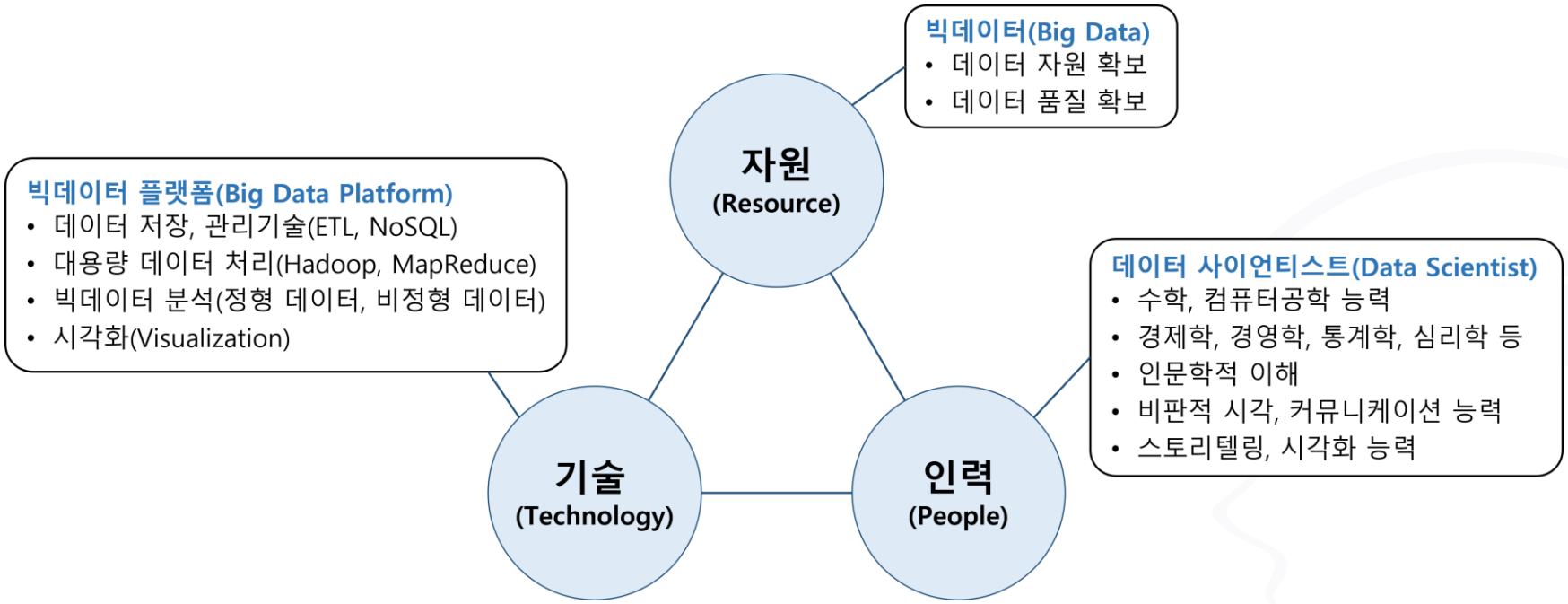
빅데이터의 활용

- **빅데이터의 활용을 위한 3요소**

구성 요소	내용
자원 (Resource) [빅데이터]	- 정형, 반정형, 비정형 데이터를 실시간으로 수집 - 수집된 데이터를 전처리 과정을 통해 품질을 향상시킴
기술 (Technology) [빅데이터 플랫홈, AI]	- 분산 파일 시스템을 통해 대용량 데이터를 분산처리 - 데이터 마이닝등을 통해 데이터를 분석 및 시각화 - 데이터를 스스로 학습, 처리할 수 있는 AI 기술을 활용
인력 (People) [알고리즈미스트, 데이터사이언티스트]	- 통계학, 수학, 컴퓨터공학, 경영학 분야의 전문 지식 확보 - 도메인 지식을 습득하여 데이터 분석 및 결과 해석

빅데이터의 활용

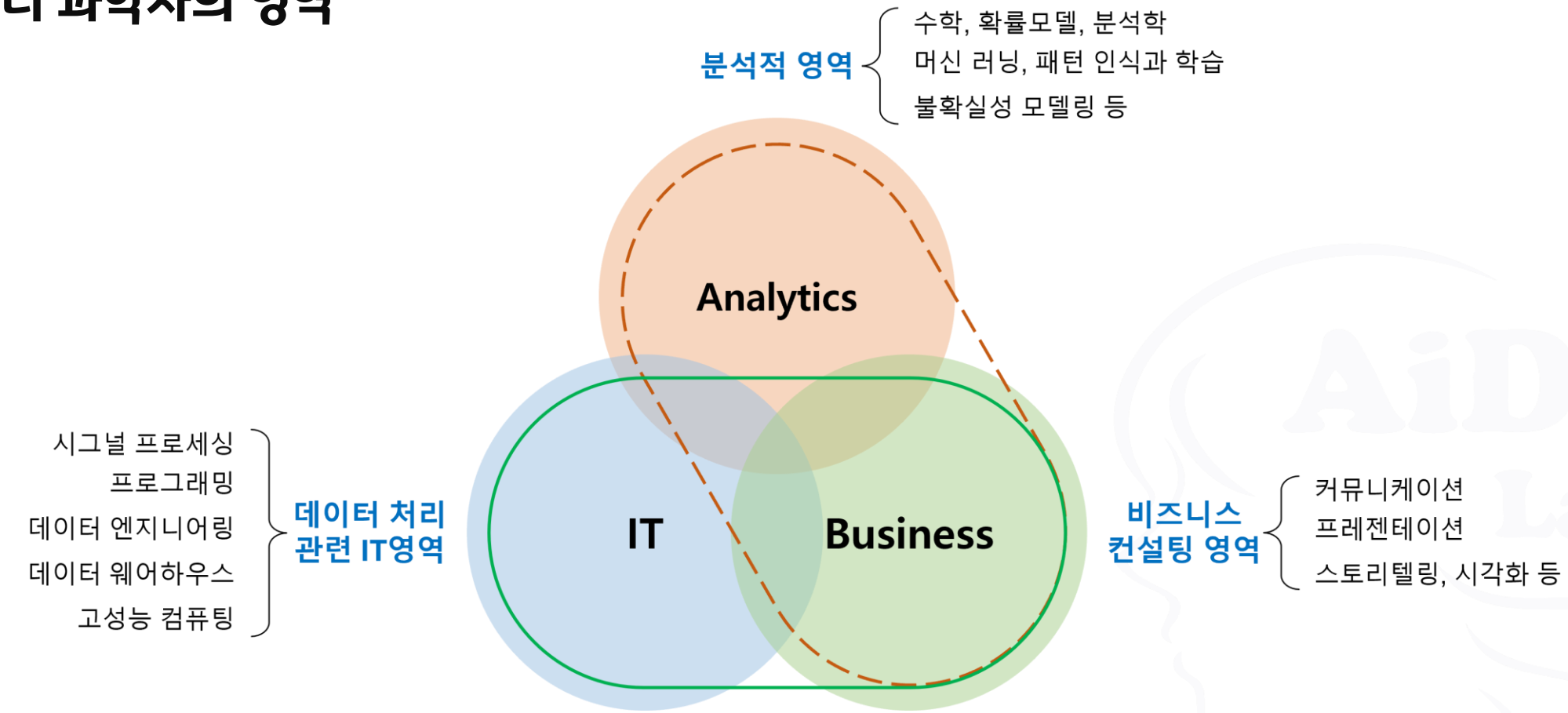
- **빅데이터의 활용을 위한 3요소**



Artificial Intelligence & Data Analysis Laboratory

빅데이터 분석에 있어서 데이터 과학자의 의의

- **데이터 과학자의 영역**



AiDA Lab.
Artificial Intelligence & Data Analysis Laboratory

빅데이터 분석에 있어서 데이터 과학자의 의의

- **데이터 과학자에게 요구되는 역량**

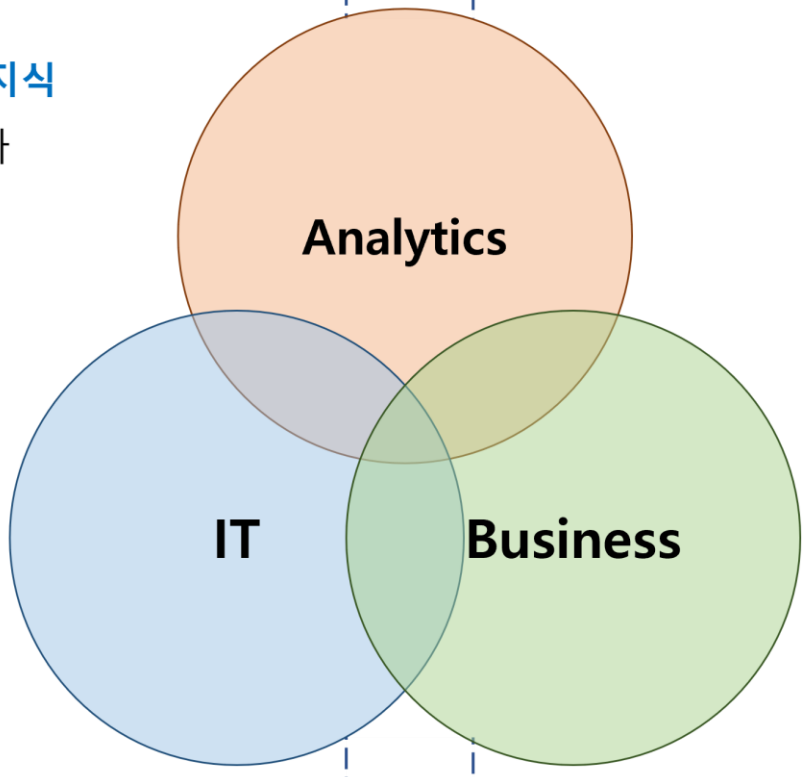
Hard Skill

빅데이터에 대한 이론적 지식

- 관련기법에 대한 이해와 방법론 습득

분석 기술에 대한 숙련

- 최적의 분석 설계 및 노하우 축적



Soft Skill

통찰력 있는 분석

- 창의적 사고, 호기심 논리적 비판

설득력 있는 전달

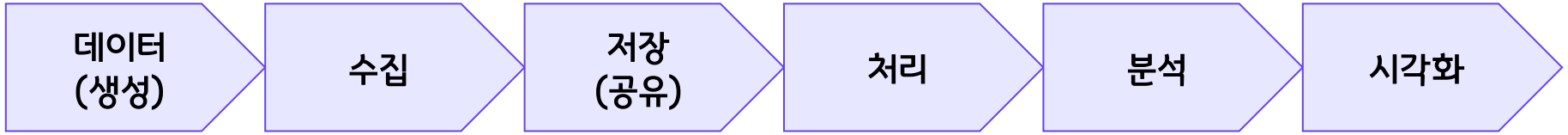
- 스토리텔링, 시각화

다분야 간 협력

- 커뮤니케이션

Artificial Intelligence & Data Analysis Laboratory

빅데이터 처리 과정과 요소 기술



- **생성**
  - 데이터베이스나 파일 관리 시스템과 같은 내부 데이터가 있음
  - 인터넷으로 연결된 외부로부터 생성된 파일이나 데이터가 있음
- **수집**
  - 크롤링을 통해 데이터 소스로부터 데이터를 검색하여 수집함
  - ETL을 통해 소스 데이터로부터 추출하고 변화하여 적재함
  - 단순한 수집이 아니라 검색 및 수집, 변환 과정을 모두 포함함
  - 로그 수집기나 센서 네트워크 및 Open API 등을 활용할 수 있음

AiDA Lab.
Artificial Intelligence & Data Analysis Laboratory

빅데이터 처리 과정

- 저장(공유)
  - 저렴한 비용으로 데이터를 쉽고 빠르게 많이 저장함
  - 정형 데이터뿐만 아니라 반정형, 비정형 데이터도 포함함
  - 병렬 DBMS나 하둡(Hadoop), noSQL 등 다양한 기술을 사용할 수 있음
  - 시스템 간의 데이터를 서로 공유할 수 있음
- 처리
  - 데이터를 효과적으로 처리하는 기술이 필요한 단계
  - 분산 병렬 및 인메모리(In-Memory) 방식으로 실시간 처리
  - 대표적으로 하둡(Hadoop)의 맵리듀스(MapReduce)를 활용할 수 있음

AiDA Lab.
Artificial Intelligence & Data Analysis Laboratory

빅데이터 처리 과정

- 분석
  - 데이터를 신속하고 정확하게 분석하여 비즈니스에 기여함
  - 특정 분야 및 목적의 특성에 맞는 분석 기법 선택이 중요함
  - 통계분석, 데이터 마이닝, 텍스트 마이닝, 기계학습 방법 등이 있음
- 시각화
  - 처리 및 분석 결과를 표, 그래프 등을 이용해 쉽게 표현하고 탐색이나 해석에 활용함
  - 정보 시각화 기술, 시각화 도구, 편집 기술, 실시간 자료 시각화 기술로 구성됨

AiDA Lab.
Artificial Intelligence & Data Analysis Laboratory

AiDA Lab.