03

비정형데이터분석

비정형 데이터의 분석 및 도구

통계·데이터과학과 장영재 교수

학습목차

1. 비정형데이터의분석
2. 분석도구의구현
3. 주요프로그래밍언어의이해(1)

01

비정형 데이터의 분석

1. 비정형 데이터의 분석

- 비정형 데이터 분석 도구는 매우 다양
  - 데이터의 원형을 계량화하는 도구, 데이터의 복잡한 구조를 효과적으로 분석할 수 있는 도구, 분석결과를 적절하게 요약하여 나타낼 수 있는 도구 등
  - 데이터의 형태와 구조의 복잡성에 기인
- 비정형 데이터 분석 도구의 의미
  - 광의의 분석 도구 : 원천 데이터로부터 데이터를 가공하여 분석이 용이한 형태로 변환하는 과정까지를 포함

    → 프로세스 전반을 아우르는 여러 기법의 집합

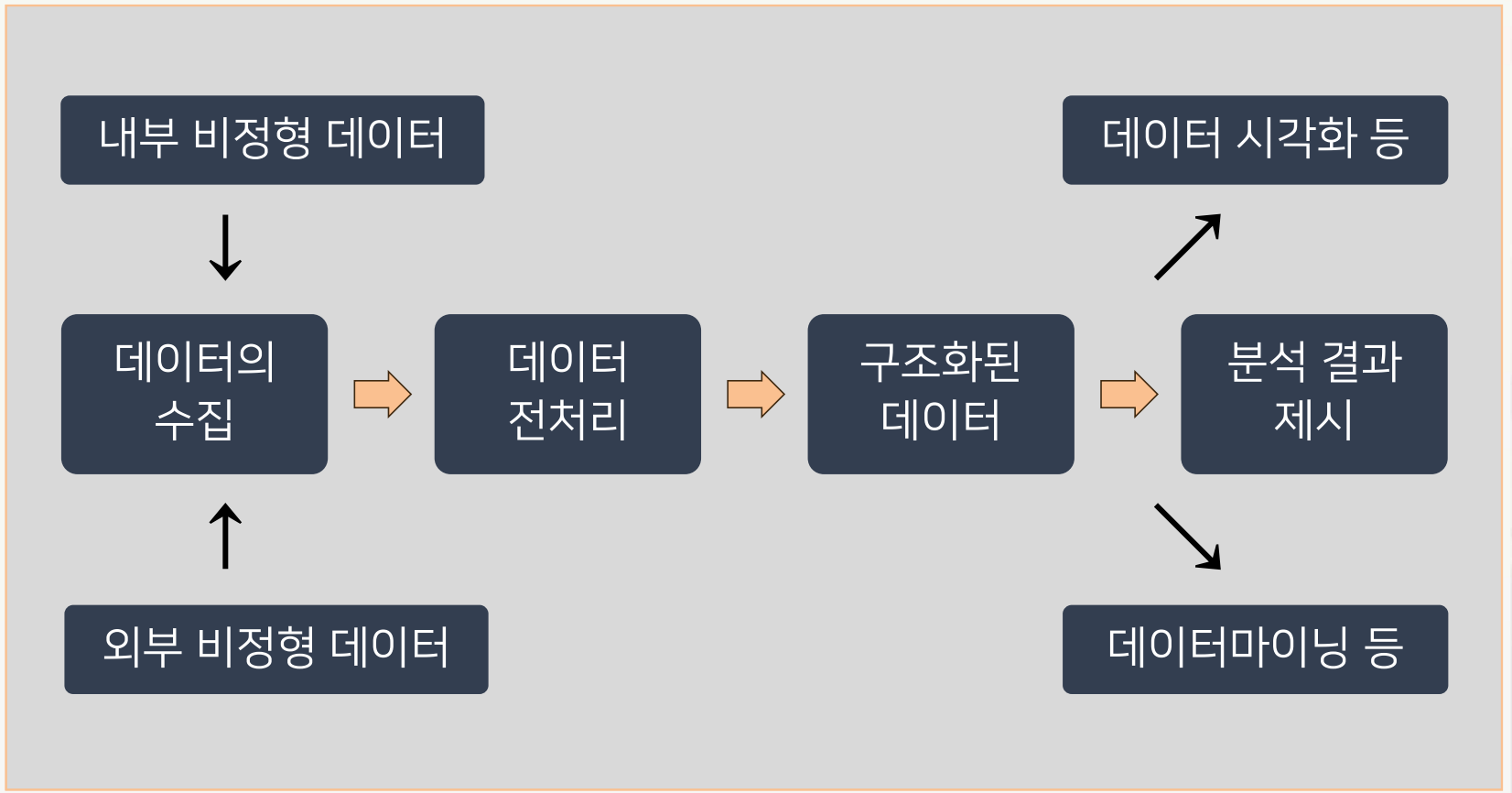
1. 비정형 데이터의 분석

- 협의의 분석 도구 : 가공 단계를 거쳐 데이터의 형태와 구조의 복잡성으로 인한 분석의 제약을 어느 정도 완화시킨 이후에 데이터를 분석하는 표준화된 도구

  → 데이터 수집이나 전처리 과정의 일부만을 포함하고 기존 정형데이터 분석에서도 널리 사용되었던 기법의 응용에 국한

> 비정형 데이터라고 하는 새로운 형태의 데이터에 기존의 도구를 어떠한 방식으로 적용할 수 있는지를 고민하고 그 연결고리에 더 큰 비중을 둠

- 비정형 데이터 분석은 데이터 전처리 단계에 투입되는 자원 투입이 상대적으로 많음
  - 최종 단계에서의 분석은 데이터마이닝 도구나 데이터시각화 도구와 같은 기존 도구 활용



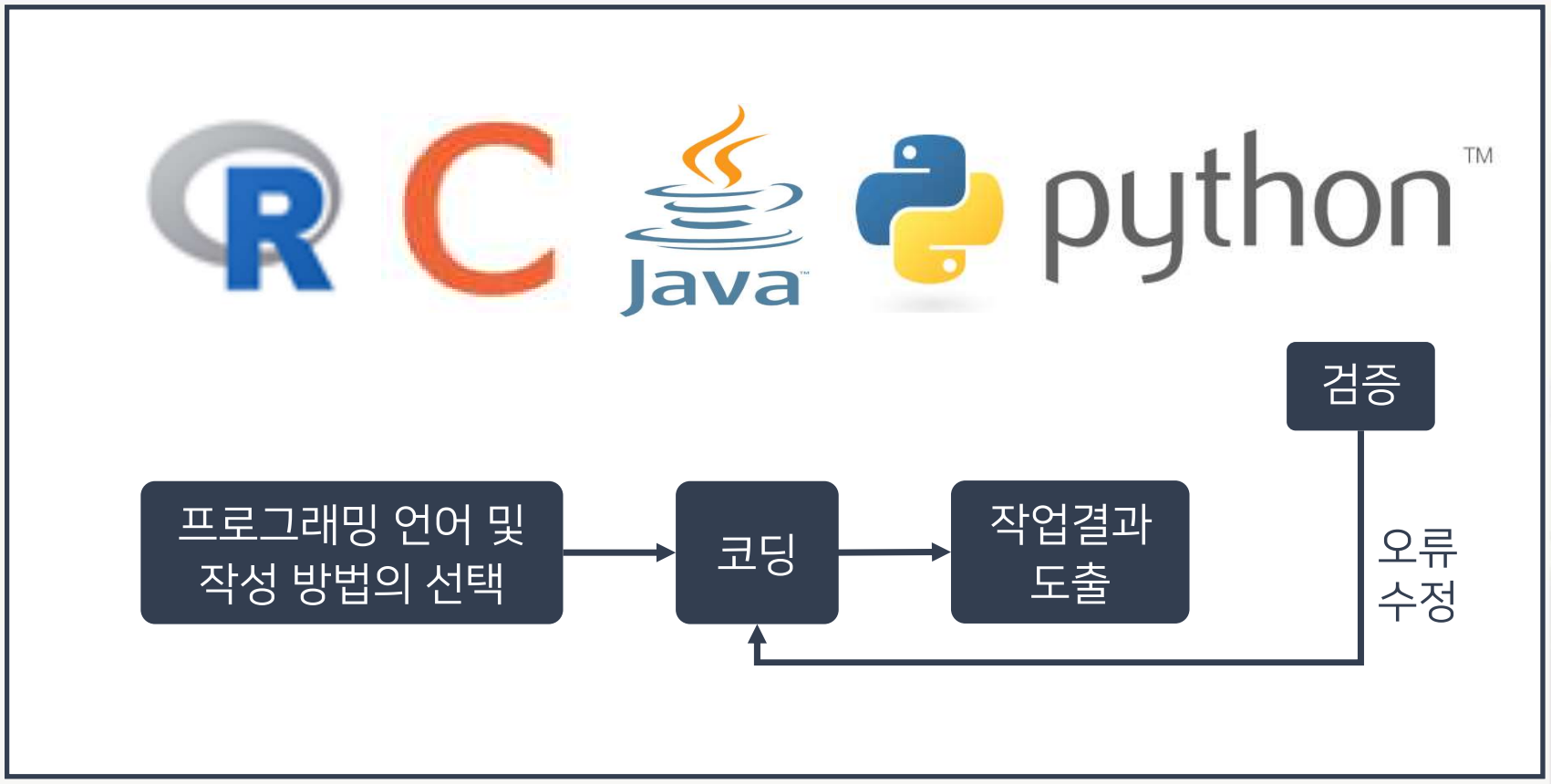
<그림> 비정형 데이터 분석 과정

02

분석도구의 구현

2. 분석도구의 구현 - ① 프로그래밍의 의의

- 비정형 데이터의 분석 도구를 개발하고 사용함에 있어서 가장 강조되는 것이 프로그래밍(programming) 능력
  - 컴퓨터 프로그래밍은 언어와 작성 방법의 선택, 코딩(coding), 에러(error) 수정 등 전반적인 절차를 지칭
    - → 코딩은 알고리즘(어떤 문제를 해결하기 위해 구성된 절차를 공식화한 것)의 구현이라는 구체적 행위를 지칭



<그림> 프로그래밍 과정

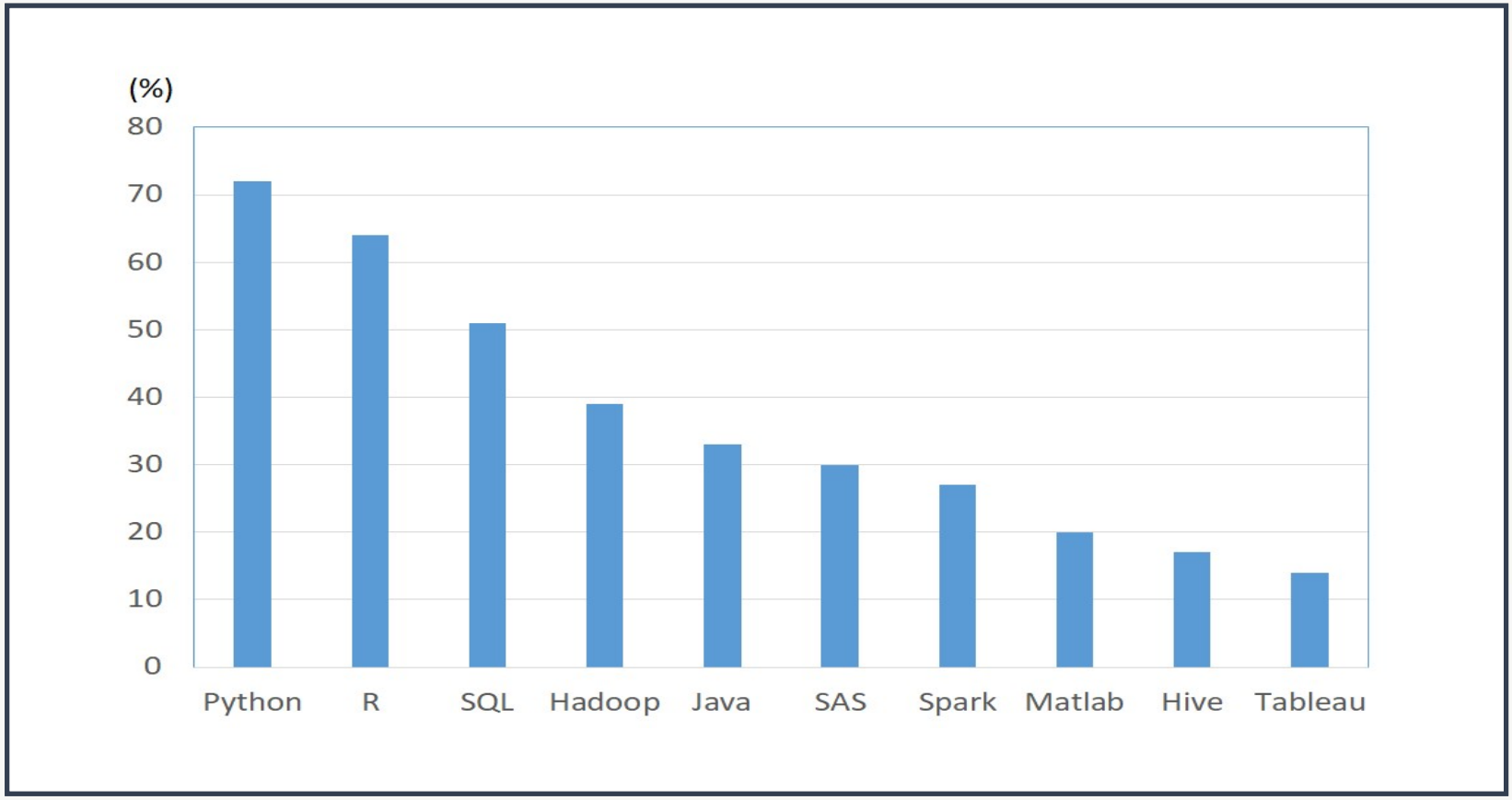
2. 분석도구의 구현 - ② 프로그래밍 언어 선택의 기준

- 위키피디아(Wikipedia)의 프로그래밍 언어 리스트에 수록된 것 만도 600개를 상회할 정도로 다양한 언어가 존재하므로 선택의 기준이 필요
  - 프로그래밍의 목적
    - → 분석의 목적에 부합하는 도구를 개발할 수 있는 언어인지 고려
  - 분석 도구의 구현을 위해 환경에 맞는 가장 적절한 언어를 선택
  - 프로그래밍 언어의 성능
    - → 효율(efficiency)이나 속도(speed)의 의미로도 이해할 수 있으며 개발자에 의존

2. 분석도구의 구현 - ③ 빅데이터 시대의 프로그래밍 언어

- 빅데이터 시대에 이르러 데이터 분석을 위한 다양한 도구들이 제안
  - 데이터과학자의 기술적 능력으로서 빅데이터를 다룰 수 있는 프로그래밍 언어에 대한 이해가 강조
  - Glassdoor라는 미국 노동시장 연구 사설 기관의 분석

→2017년 1월부터 7월까지 자사의 구인 사이트에 올라가 있는 데이터 과학 구인 광고 중 업무에 필요한 기술(skill)로 적시되어 있는 프로그래밍 언어를 조사



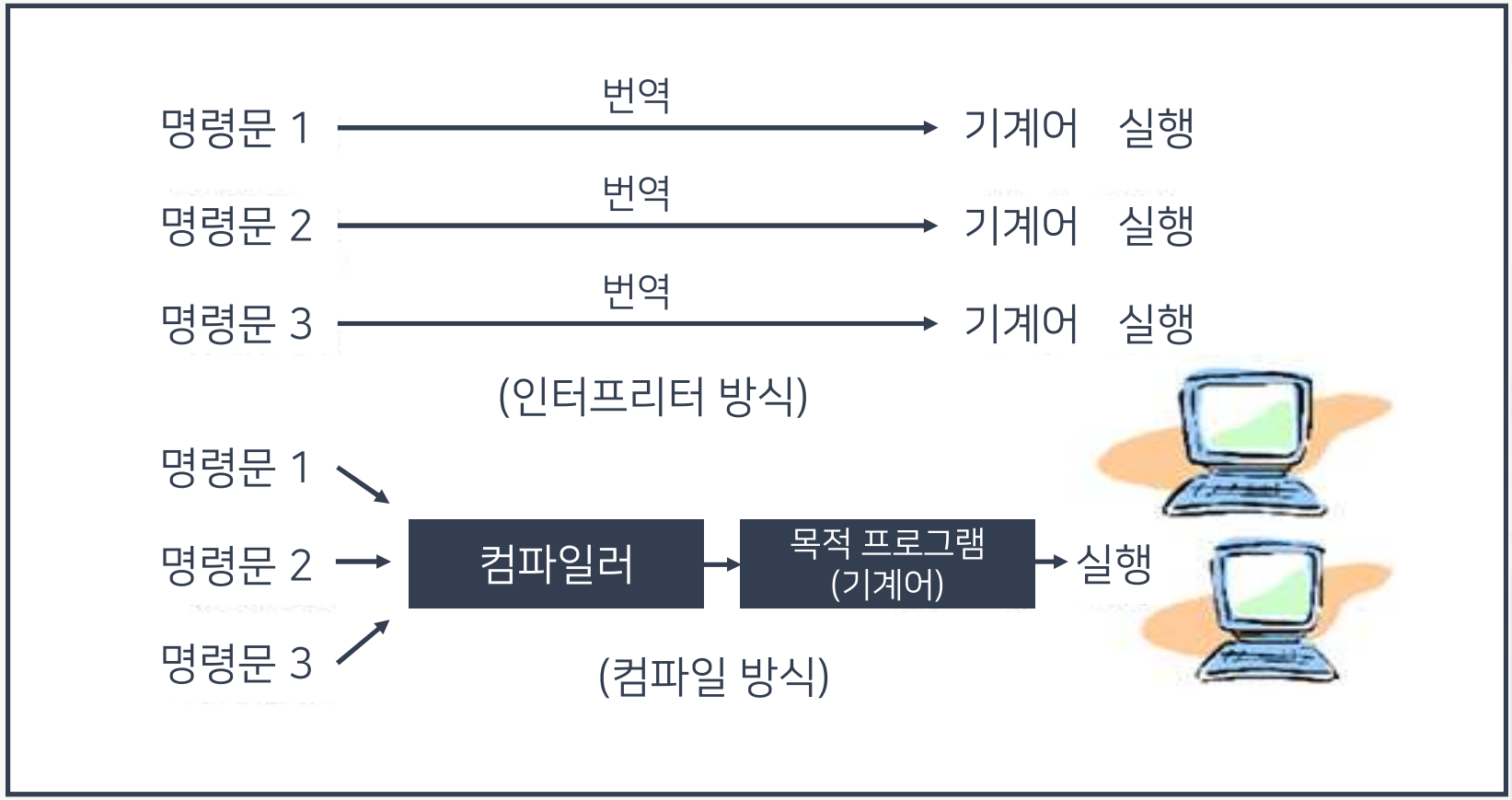
<그림> 2017년 구인광고 중 언급된 주요 프로그래밍 언어(출처: Glassdoor)

03

주요 프로그래밍 언어의 이해(1)

3. 주요 프로그래밍 언어의 이해 - ① 파이썬의 개요

- Guido van Rossum이 발표한 고급 프로그래밍 언어
  - 자연어에 가까워 직관적으로 이해하기 쉬운 언어
    - → 코딩 작업도 간결하게 할 수 있으며 문법도 쉬우며 다른 프로그래밍 언어와 호환성이 높은 편이며 풍부한 라이브러리가 특징
    - → 인터프리터(interpreter) 언어로서 프로그래밍 언어의 소스 코드(프로그래밍 언어로 작성한 프로그램 코드)를 입력하여 컴파일을 거치지 않고 실행



<그림> 인터프리터와 컴파일 방식의 비교

3. 주요 프로그래밍 언어의 이해 - ① 파이썬의 개요

- 오픈소스(open source)이며 무료로 사용이 가능하다는 점과 객체 지향성이 특징
- 동적타이핑(dynamic typing), 즉 자료의 형태를 미리 명시적으로 지정해주지 않고 인터프리터에 그 처리를 담당하게 하는 방식을 따름

3. 주요 프로그래밍 언어의 이해 - ① 파이썬의 개요

- 객체지향 프로그래밍 언어(object oriented programming language)란 클래스(class), 객체(object), 기능(method) 등의 요소를 지니고 있는 프로그래밍 언어
  - 프로그래밍을 여러 개의 독립된 단위인 객체들의 조합으로 간주하는 시각이 반영

→ 객체는 속성이나 동작을 지닌 물건으로 이해할 수 있음

> 특정 스마트폰 : 속성으로는 제조사, 제조사, 모델, 디자인, 가격 등을 꼽을 수 있으며 동작으로는 전화통화, 문자송수신, 검색, 계산 등 다양한 기능을 꼽을 수 있음

→ 클래스는 이러한 모든 객체를 아우르는 설계도의 개념

> 스마트폰이라는 하나의 클래스가 정의되면 제조사, 모델, 디자인, 가격 및 기능에 따라 여러 개의 객체를 생성

3. 주요 프로그래밍 언어의 이해 - ② R의 개요

- R은 뉴질랜드 오클랜드(Auckland) 대학의 로스 이하카(Ross Ihaka)와 로버트 젠틀맨(Robert Gentleman) 교수의 주도하에 만들어진 프로그래밍 언어로서 많은 장점을 지님
  - 매우 다양한 통계적 분석과 우수하고 다양한 그래픽 방법을 제공
  - 통계학 분야 뿐만 아니라 금융, 생명공학, 조사, 지리정보 등 여러 분야의 이용자들도 R로 만들어진 응용 프로그램을 이용
  - 오픈소스 무료프로그램, 인터프리터 언어, CRAN에 탑재된 수많은 패키지를 이용할 수 있다는 점 등 파이썬과 유사한 특징을 지님
    - → R은 통계적 분석을 위해 최적화된 프로그램 언어, 파이썬은 일반적인 데이터과학용 프로그래밍 언어라고 구분해 볼 수도 있음

3. 주요 프로그래밍 언어의 이해 - ② R의 개요

- 다만 데이터를 물리적 메모리에 저장하고 작업을 수행하므로 데이터의 크기가 커질수록 연산속도가 크게 저하되는 문제가 발생하는 등 단점도 있음
  - 하드웨어 자체의 발전, 특히 컴퓨터에 장착하는 메모리 용량의 증가로 인해 이러한 제약은 많이 해소되고 있는 상황

다음시간안내

04

주요 프로그래밍 언어의 이해