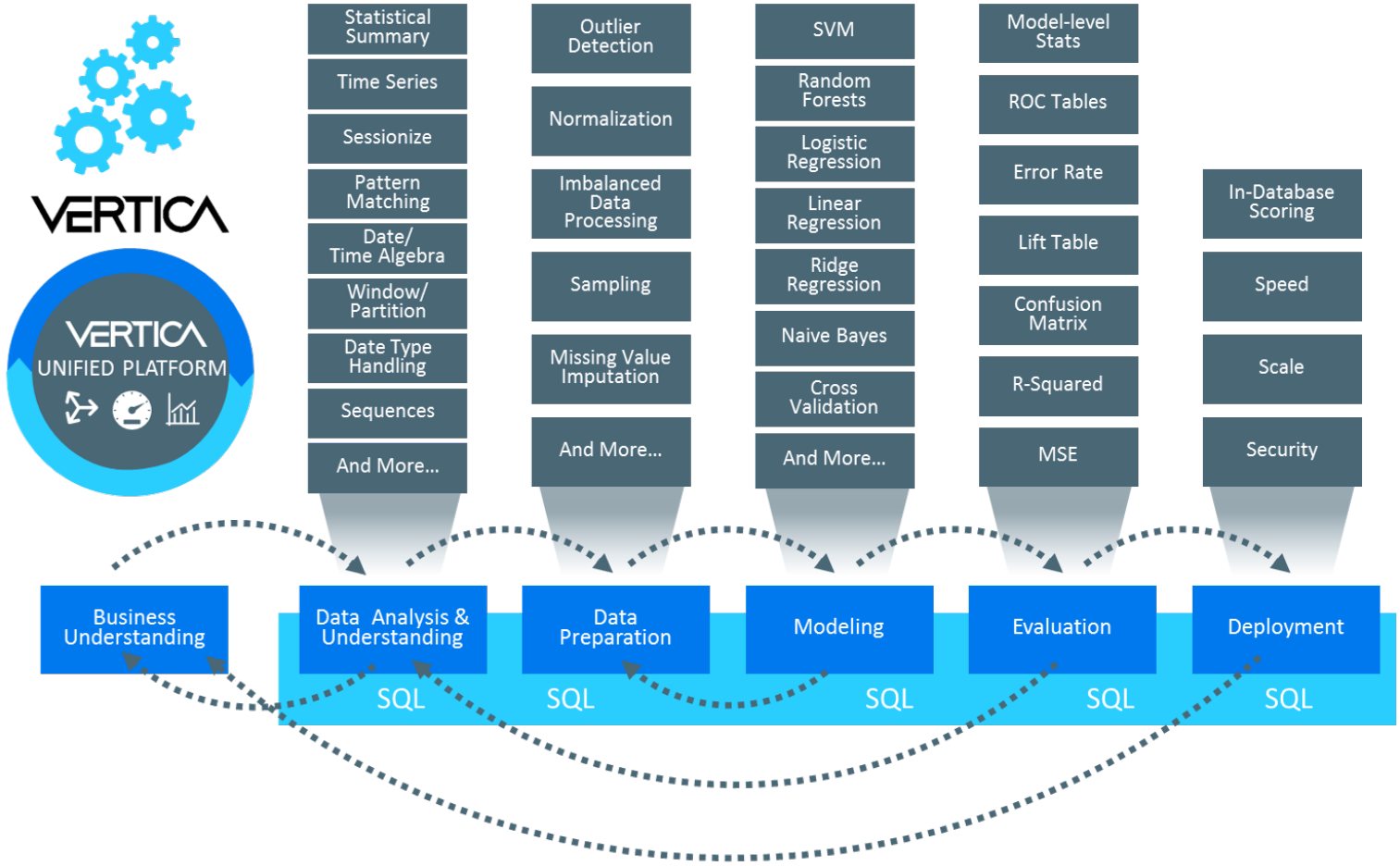


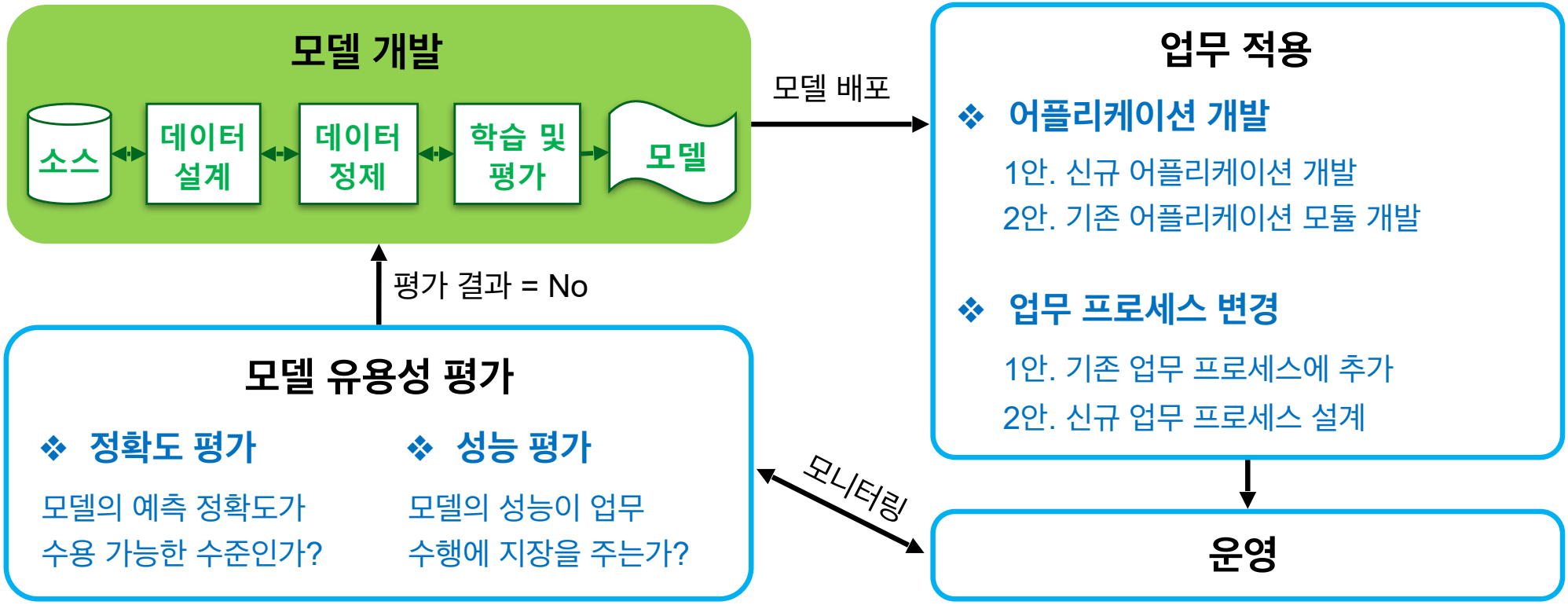
Vertica in-DB Machine Learning 데모

Softline R&D센터

Machine Learning 라이프 사이클

머신러닝 시스템은 기존에 매뉴얼하게 진행되었거나, 사전 정의된 로직에 의해 수행되었던 프로세스를 머신러닝 기반 모델을 통해 자동화하는 것. 핵심은 모델 개발이며, 이는 일반 어플리케이션 개발과 근본적으로 다른 관점에서 접근해야 함.

- 머신러닝의 장점은 모델을 빠르고 정확하게 만들어 낼 수 있다는 점.
- 이 장점을 살려 정기적 모니터링 / 평가를 통해 모델을 꾸준히 개선하며 활용 (Agile / Iterative 방법론).
- 개발의 핵심은 정확한 모델을 얼마나 빠르게 만들어 내는가에 있음.



SoftLine AI Big Data Cognitive Business R&D Center

Machine Learning 방법론

관련 업무에 대한 이해를 바탕으로 해결하고자 하는 문제를 정확히 정의하고, 문제해결을 위해 필요한 데이터를 설계한 후, 필요한 데이터를 다양한 원천으로부터 확보 및 정제하는 것이 가장 중요함. 이후 모델을 생성하고 운영 과정에서 모니터링 하면서 업그레이드 및 재개발함. 이 모든 과정은 구간 혹은 전체 반복을 전제로 함.

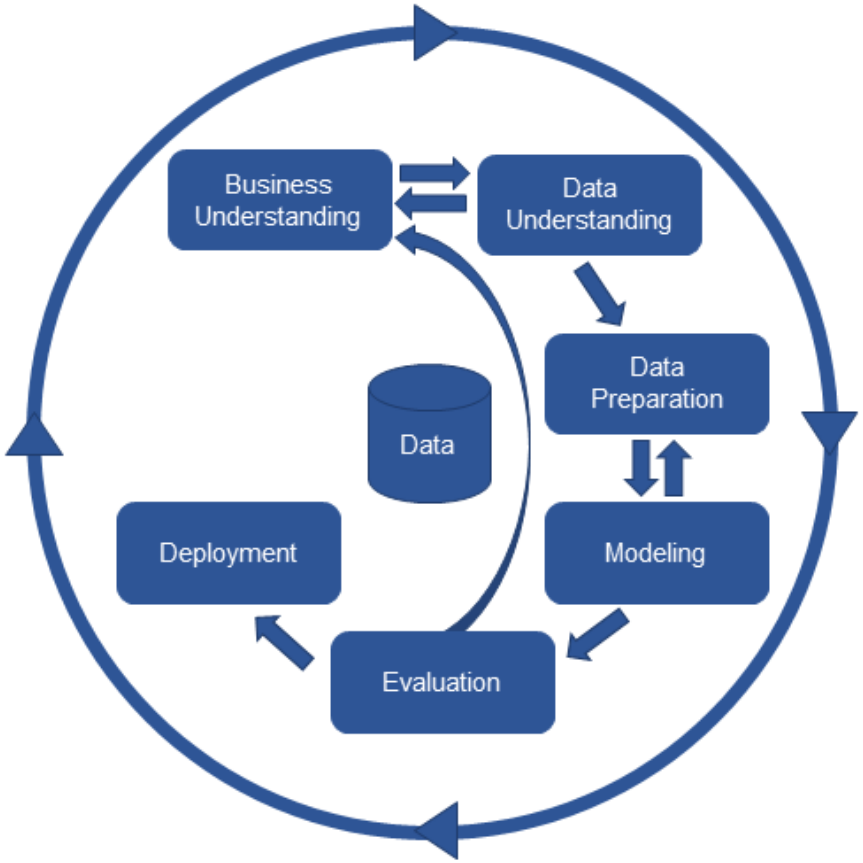
투입자원비중

80%

1. **업무 이해 (Business understanding)**
   프로젝트를 통해 해결해야 하는 업무문제를 완벽하게 이해하고, 업무 개념을 분류하고 정의함.
2. **데이터 이해 (Data understanding)**
   업무문제에 대한 이해를 바탕으로 필요한 데이터를 정의하고, 활용 가능한 데이터에 대한 이해.
3. **데이터 준비 (Data preparation)**
   머신러닝 알고리즘을 훈련시키기 위한 데이터를 ABT (Analytics Base Table) 이라고 함. 데이터 준비 단계는 바로 이 ABT를 설계하고 구현.

20%

4. **모델링 (Modeling)**
   후보 모델군을 선정하고, 이 중 최적의 모델을 선정하는 단계
5. **평가 (Evaluation)**
   선정된 모델 배포 이전에 업무문제 해결에 실질적인 도움이 되는지 평가
6. **배포 (Deployment)**
   평가를 마친 예측 모델을 조직 내 프로세스에 성공적으로 안착시키기 위한 단계

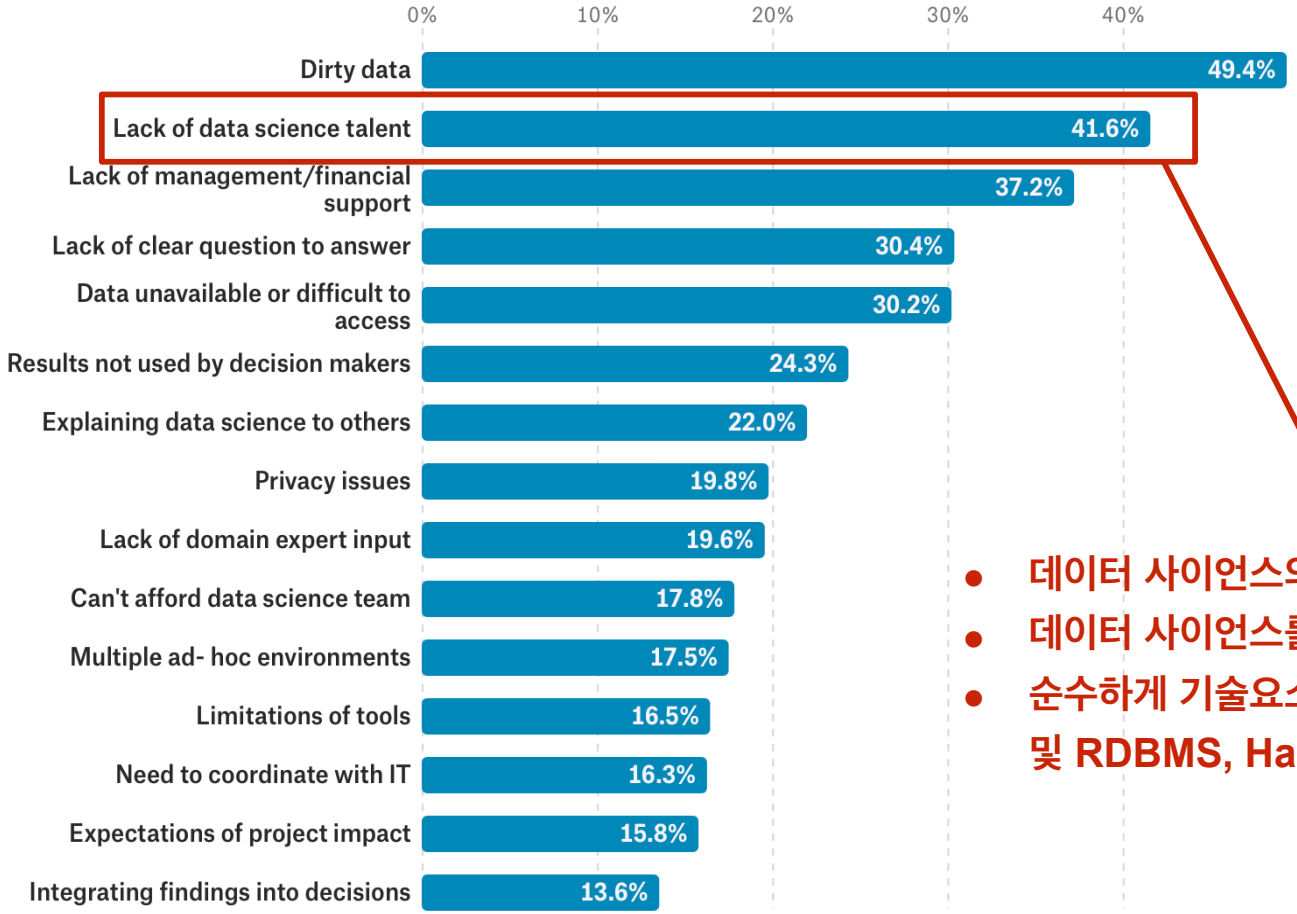


SoftLine A.I Big Data Cognitive Business R&D Center

The State of ML and Data Science (Kaggle Survey)

2017년 Kaggle에서 16,000명의 데이터 사이언티스트 및 관련 종사자를 대상으로 수행한 설문조사 결과를 통해 몇 가지 시사점을 배울 수 있음. https://www.kaggle.com/surveys/2017

1. 데이터 사이언스의 가장 큰 장애물



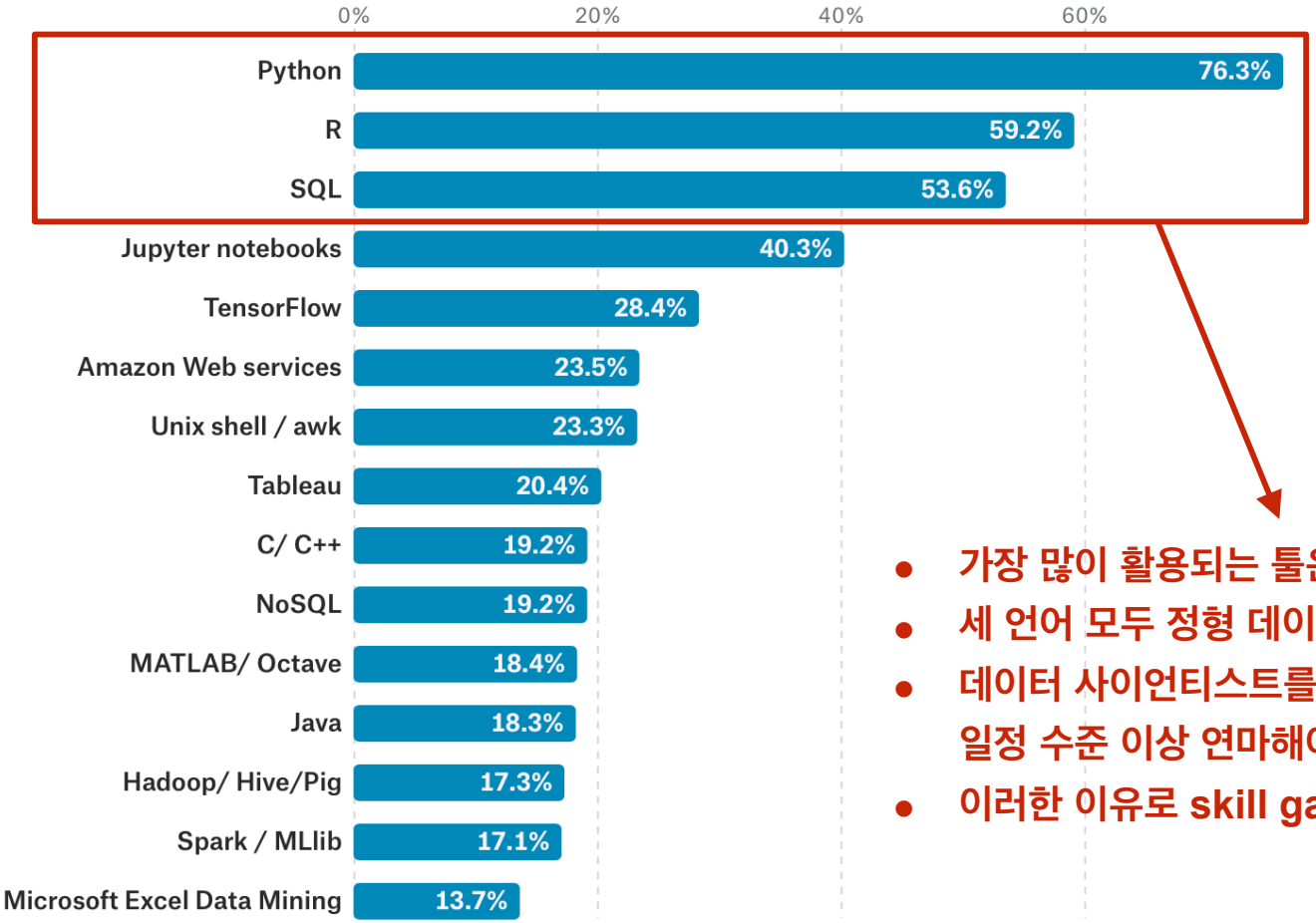
- 데이터 사이언스의 두번째 장애물은 인력 부족.
- 데이터 사이언스를 위한 수학 / 통계관련 지식 및 업무지식 포함.
- 순수하게 기술요소만 고려하더라도 R, Python, SQL 등의 언어 및 RDBMS, Hadoop, Spark 등 플랫폼에 대한 이해도도 요구됨.

SoftLine AI Big Data Cognitive Business R&D Center

The State of ML and Data Science (Kaggle Survey)

2017년 Kaggle에서 16,000명의 데이터 사이언티스트 및 관련 종사자를 대상으로 수행한 설문조사 결과를 통해 몇 가지 시사점을 배울 수 있음. https://www.kaggle.com/surveys/2017

2. 데이터 사이언스 업무에 활용하는 툴



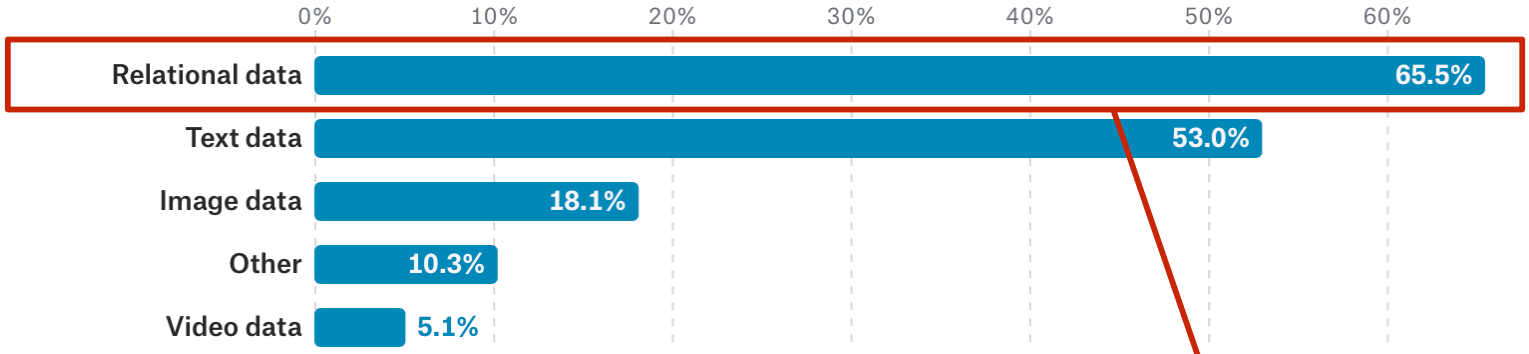
- 가장 많이 활용되는 툴은 Python, R, SQL
- 세 언어 모두 정형 데이터를 다루는 데에 특화되어 있음
- 데이터 사이언티스트를 지망하는 인력들은 이처럼 다양한 툴들을 일정 수준 이상 연마해야 함
- 이러한 이유로 skill gap이 AI/ML에 있어 큰 장애물이 됨

SoftLine AI Big Data Cognitive Business R&D Center

The State of ML and Data Science (Kaggle Survey)

2017년 Kaggle에서 16,000명의 데이터 사이언티스트 및 관련 종사자를 대상으로 수행한 설문조사 결과를 통해 몇 가지 시사점을 배울 수 있음. https://www.kaggle.com/surveys/2017

3. 가장 많이 활용하는 데이터

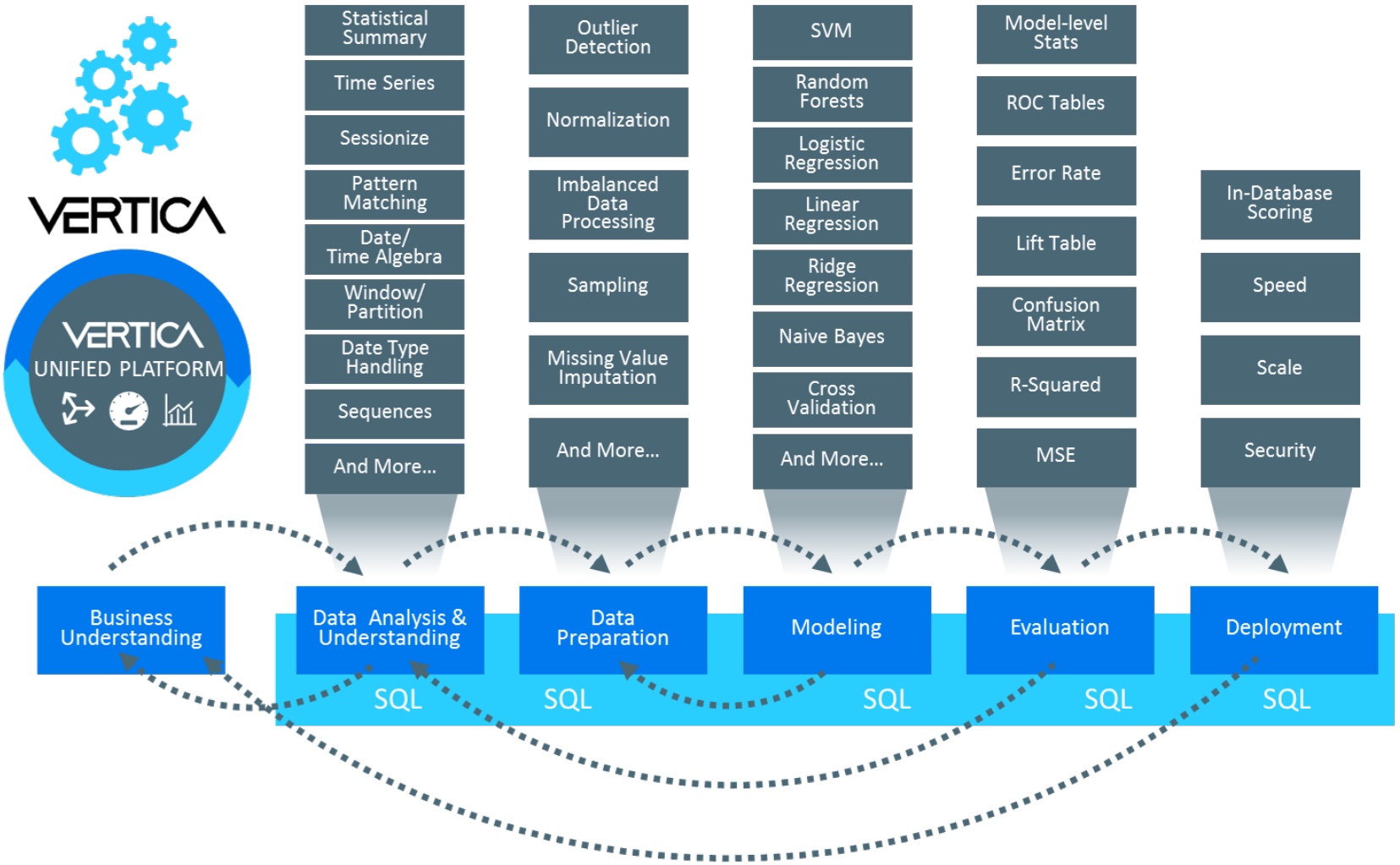


- **AI / ML을 포함한 모든 데이터 분석은 분석 가능한 데이터를 대상으로 이루어짐**
- **이미지, 텍스트 등 비정형 데이터도 분석 가능한 정형 데이터 형태로 전환된 후 분석됨**
- **따라서, 과거이든 지금이든 앞으로든 정형 데이터가 핵심**

SoftLine AI Big Data Cognitive Business R&D Center

Vertica in-DB Machine Learning

머신러닝 방법론 상의 모든 단계를 in-DB로 구현할 수 있음. 즉, 새로운 제품이나 기술을 습득하지 않고도 보편적으로 사용되는 **ANSI-SQL 만으로 훈련 데이터 생성에서부터 모델 학습, 실행까지 하나의 플랫폼에서 구현 가능함**.

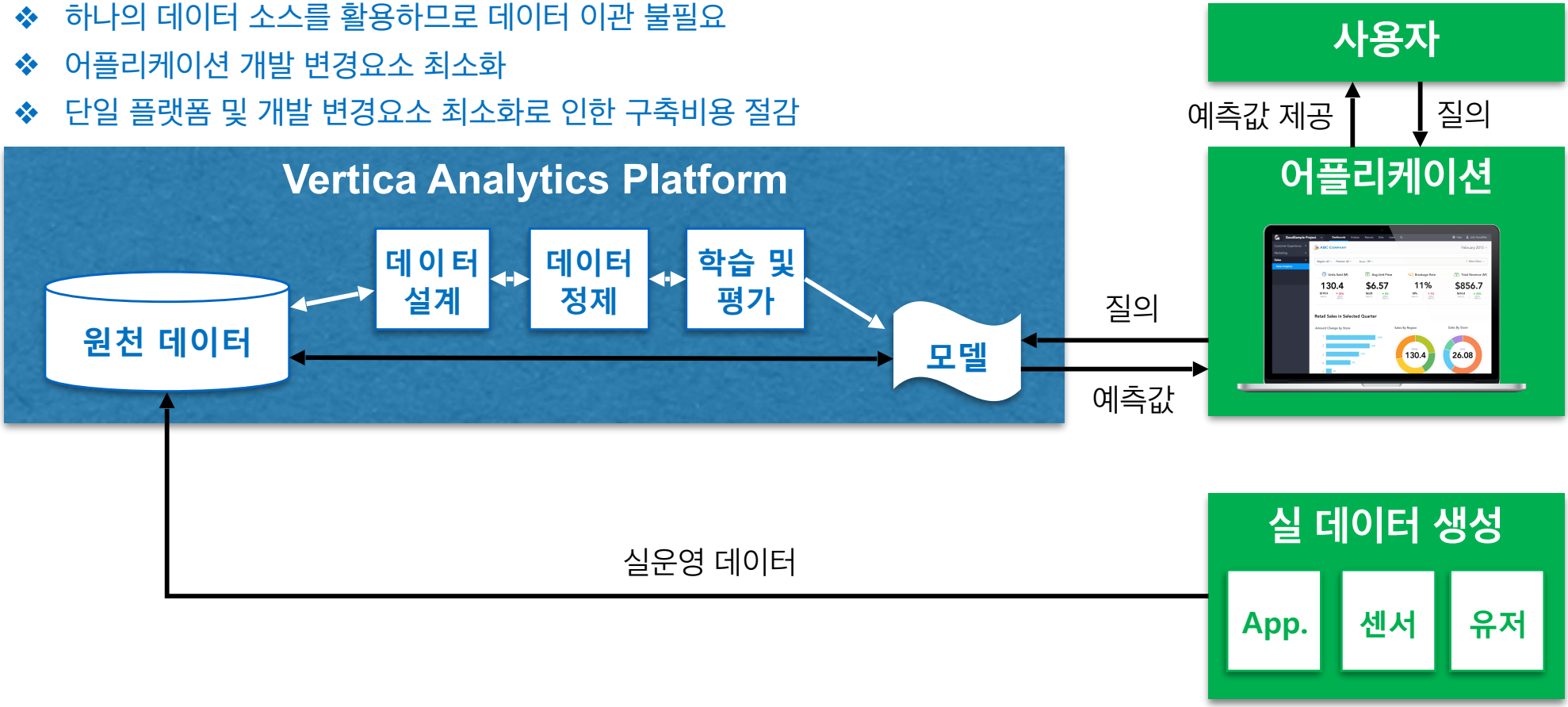


SoftLine A.I Big Data Cognitive Business R&D Center

Vertica in-DB Machine Learning

Vertica Analytics Platform을 활용한 on-premise 머신러닝 플랫폼 구성방안. **ANSI-SQL 만으로 훈련 데이터 생성에서 부터 모델 학습, 실행까지 하나의 플랫폼에서 구현 가능함**.

- 단일 플랫폼에서 머신러닝 예측분석 모델 개발에서부터 실행까지 처리
- SQL만으로 머신러닝 예측분석에 필요한 모든 프로세스 처리
- 하나의 데이터 소스를 활용하므로 데이터 이관 불필요
- 어플리케이션 개발 변경요소 최소화
- 단일 플랫폼 및 개발 변경요소 최소화로 인한 구축비용 절감

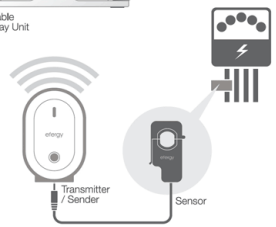


SoftLine AI Big Data Cognitive Business R&D Center

데모 시나리오

아일랜드 전역의 가구 및 기업에 배치된 전력 사용량 측정 데이터와 (15분 간격) 해당 지역의 날씨 데이터를 (30분 간격) 기초 데이터로 함. 각 센서의 위치, 센서가 설치된 장소의 종류에 대한 데이터도 포함.

http://www.ucd.ie/issda/data/commissionforenergyregulationcer/



전력 사용량 센서

날씨 센서

1. 예측모델 개발

센서 측정 데이터 수집

VERTICA

모델 생성 및 평가

VERTICA

데이터 정제

Balancing

Normalizing

Imputing

Outliers

VERTICA

2. 예측모델 운영

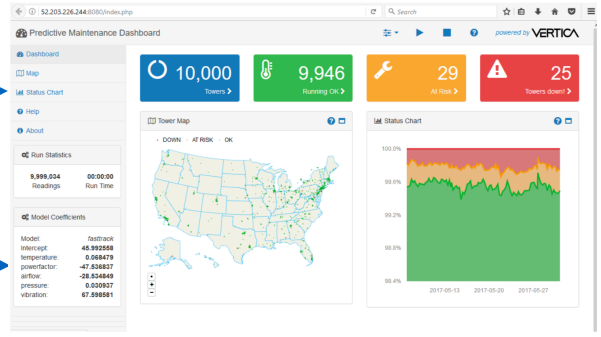
Micro Batch

센서 측정 데이터 수집
(현재 실시간 30초 간격)

VERTICA

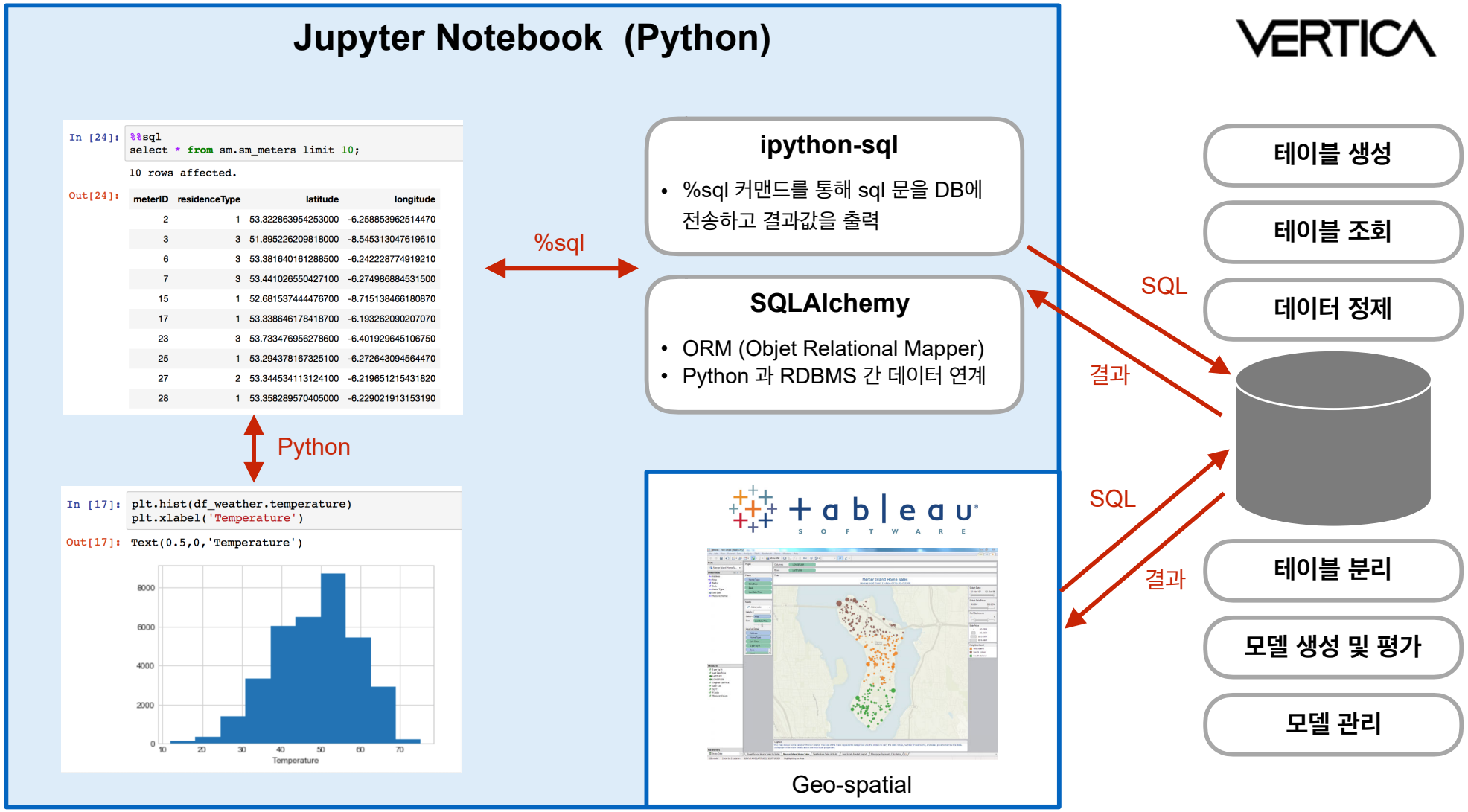
예측 모델
(장애 발생 사전 예측)

VERTICA



SoftLine A.I Big Data Cognitive Business R&D Center

데모 시스템 개념도



SoftLine AI Big Data Cognitive Business R&D Center

Demo