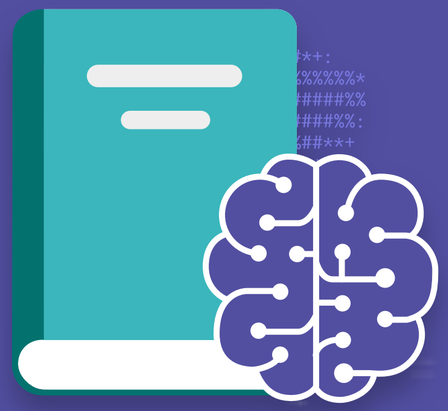
/* elice */

문과생을 위한 머신러닝

2주차: 데이터 과학자 이해하기



David Oh 선생님

목차

1. 데이터 과학자 Data Scientist

데이터 과학자 ?

데이터 과학은 컴퓨터를 활용해서

데이터를 분석하고 현실의 문제들을 해결하는 것

Data Scientist:

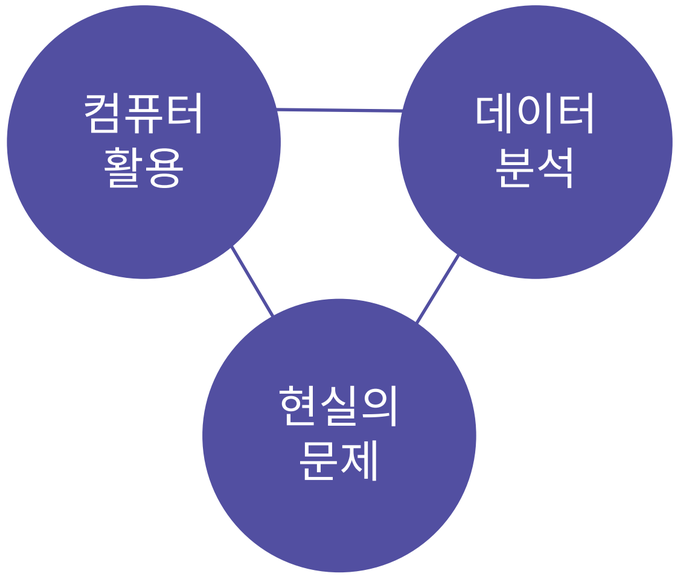
The Sexiest Job of the 21st Century

Meet the people who can coax treasure out of messy, unstructured data.

by Thomas H. Davenport and D.J. Patil

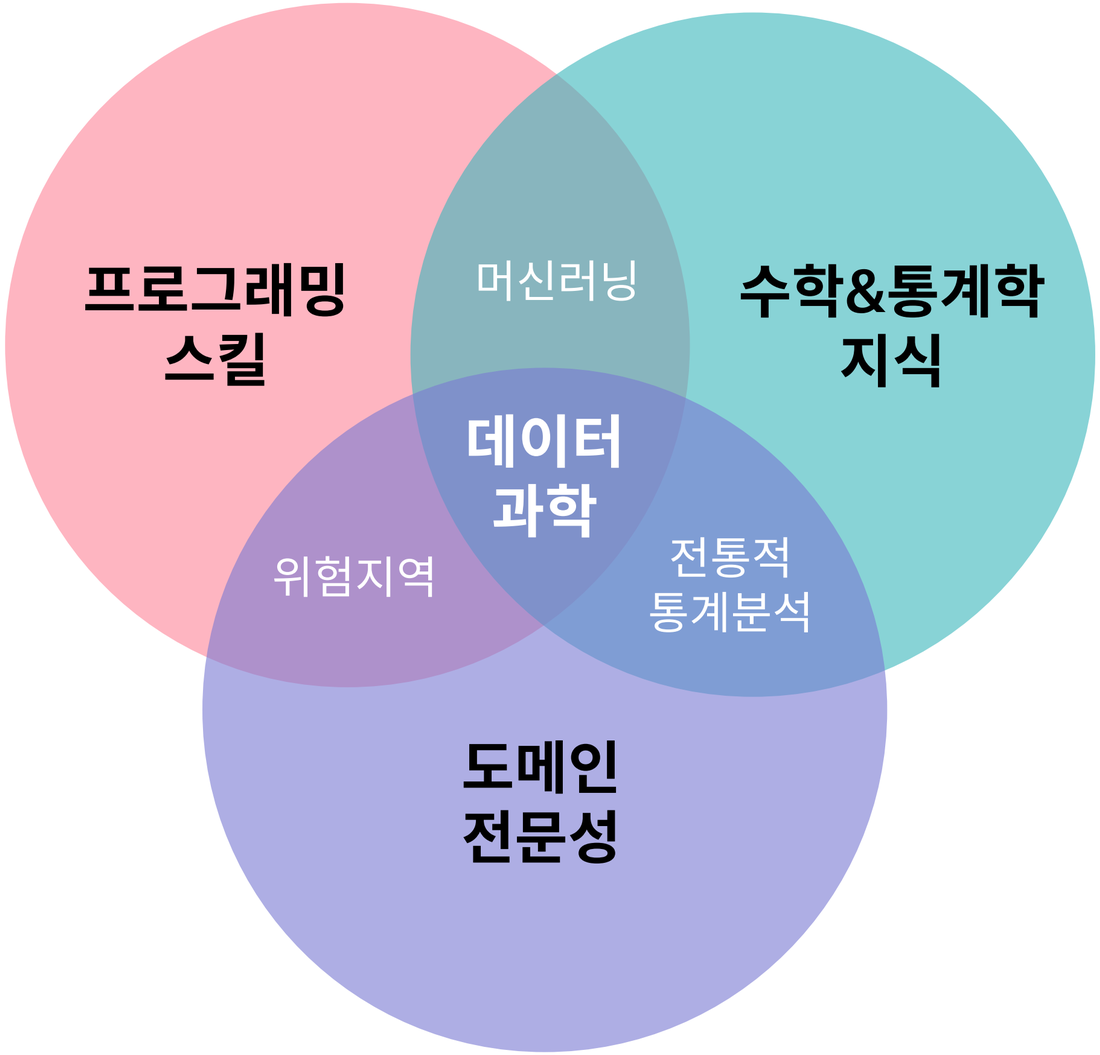
When Jonathan Goldman arrived for work in June 2006 at LinkedIn, the business networking site, the place still felt like a start-up. The company had just under 8 million accounts, and the number was growing quickly as existing members invited their friends and colleagues to join. But users weren't seeking out connections with the people who were already on the site at the rate executives had expected. Something was apparently missing in the social experience. As one LinkedIn manager put it, "It was like arriving at a conference reception and realizing you don't know anyone. So you just stand in the corner sipping your drink—and you probably leave early."

70 Harvard Business Review October 2012



Referenced Harvard Business Review 2012 Oct.

데이터 과학자에게 요구되는 실무능력은?



Adapted from Drew Conway's Data Science Venn Diagram

흔히들 말하는 데이터 과학자 Skill Sets

Programming

- 컴퓨터 과학 지식
- 프로그래밍 언어 (Python/R)
- 데이터베이스 언어(SQL/NoSQL)
- Relational Algebra
- 병렬 처리 컴퓨팅
- MapReduce 개념
- Hadoop/Hive/Pig
- AWS 같은 플랫폼 사용 경험

Math & Statistics

- 연구 계획 (Experiment design)
- Machine Learning
- Statistical modeling
- 베이지안 추론
- 선형대수, 미적분
- Supervised Learning
- Unsupervised Learning
- Optimization

Domain Knowledge

- 비즈니스 이해/지식
- Collaborative
- 데이터에 대한 호기심
- 전략적 사고/기획력
- 문제 해결능력
- Proactive/Creativity

Communication

- 상급자와의 원활한 의사소통 능력
- 스토리텔링 능력
- 데이터 기반 인사이트를 의사결정에 활용하는 능력
- ppt, doc 등 문서 작성 능력
- 시각화(Visualization)
- 발표/설득력

데이터 과학자 !

데이터 과학은 컴퓨터를 활용해서

데이터를 분석하고 현실의 문제들을 해결하는 것


Data Scientist:
The Sexiest Job of the 21st Century

컴퓨터와 IT기술을 활용하고 프로그래밍을 할 수 있는 능력을 가진 사람이

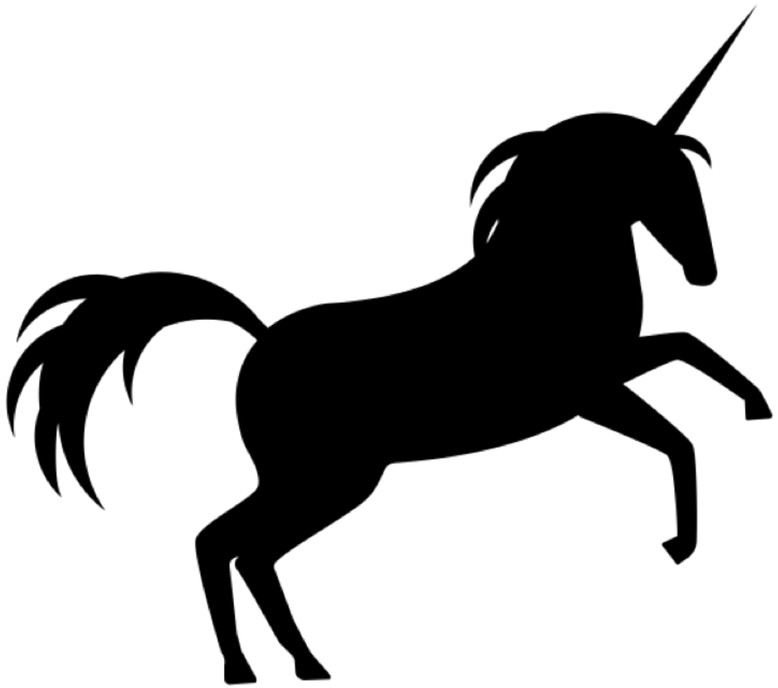
수학과 통계학 지식을 이용해서 **도메인의 문제를 해결**하는 사람

ing in the social experience. As one LinkedIn manager put it, "It was like arriving at a conference reception and realizing you don't know anyone. So you just stand in the corner sipping your drink—and you probably leave early."

70 Harvard Business Review October 2012

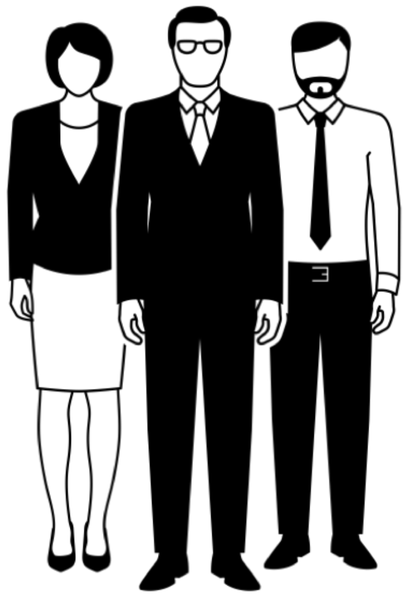
Referenced Harvard Business Review 2012 Oct.

왜 유니콘이라고 부를까?



혼자서 다 못해

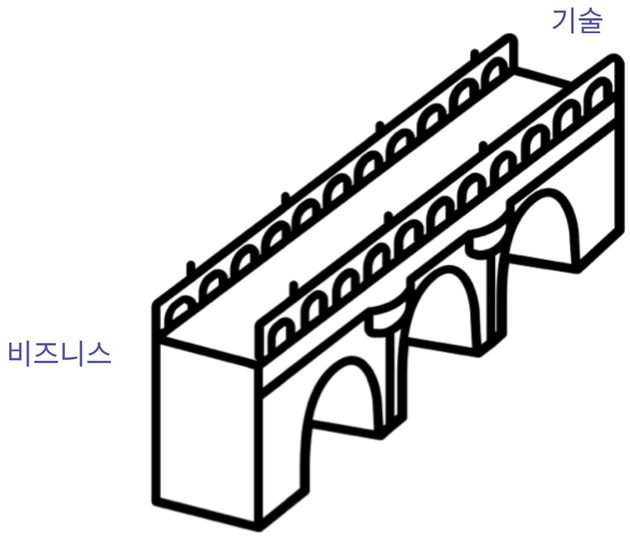
Team Sports



그래서 중요한 건 협업

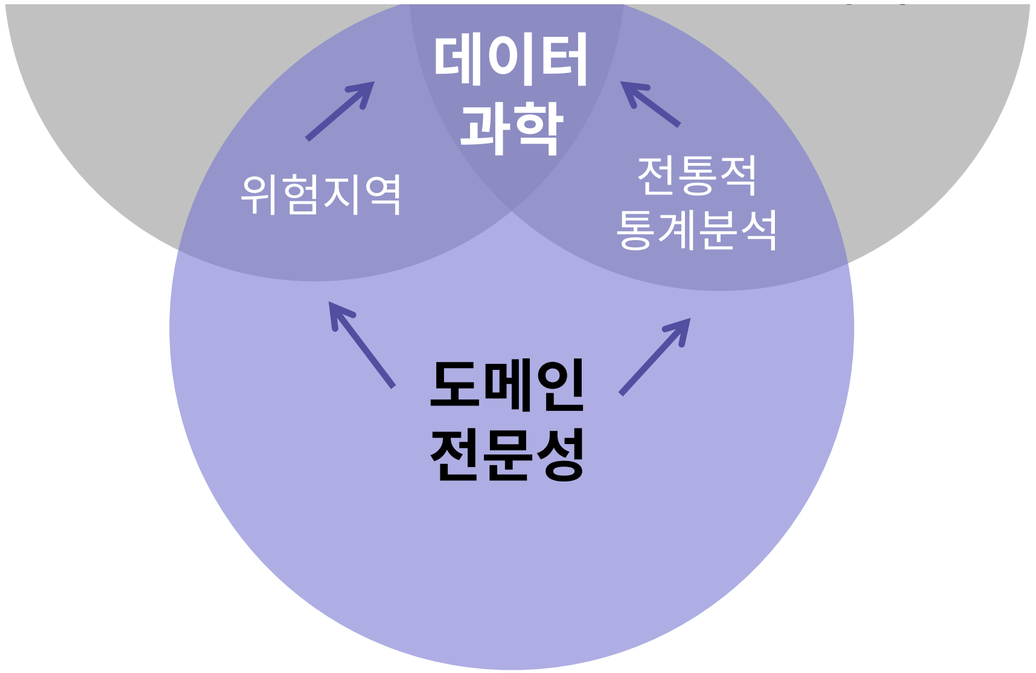
의사소통

가교 역할



문과생 데이터 과학자의 방향

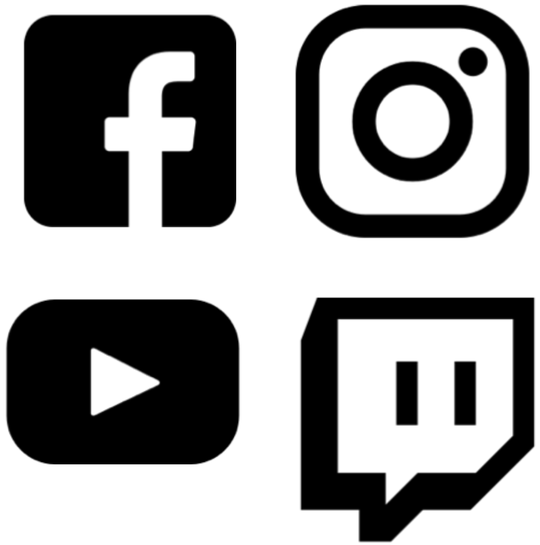
It first starts by providing you can do something,
that you can make something. - DJ Patil



Adapted from Drew Conway's Data Science Venn Diagram

도메인 전문성의 중요성 (1)

SNS

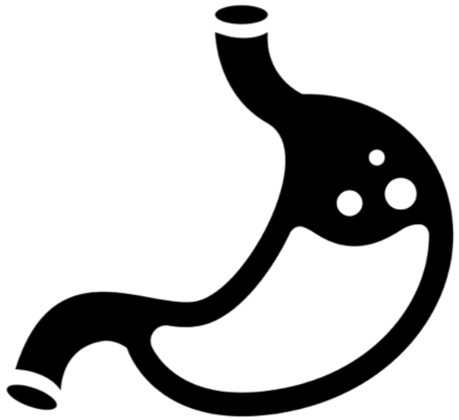


AU, CTR

도메인 전문성의 중요성 (2)

NSAID

Gastrointestinal Bleed

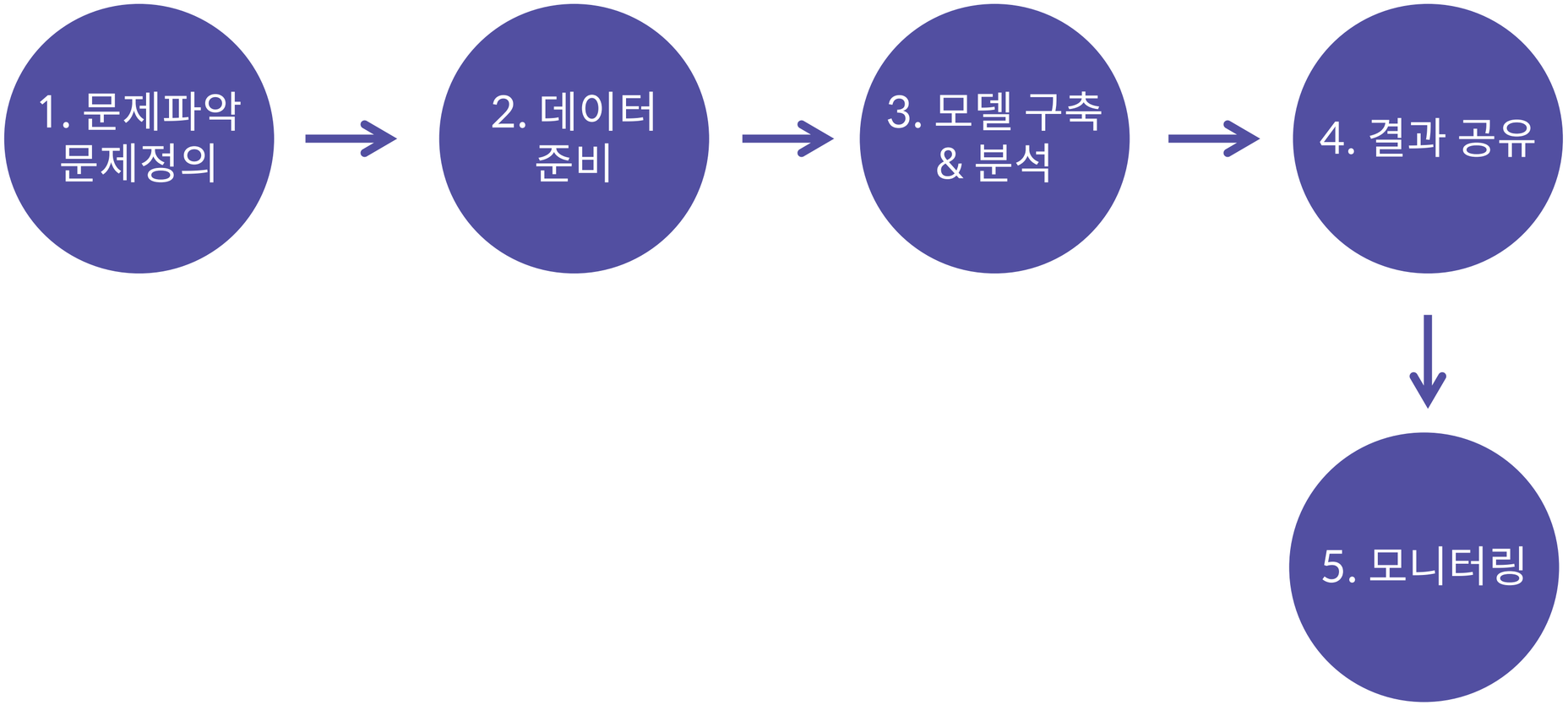


2. 심장질환을 앓는 사람은 누구 일까?

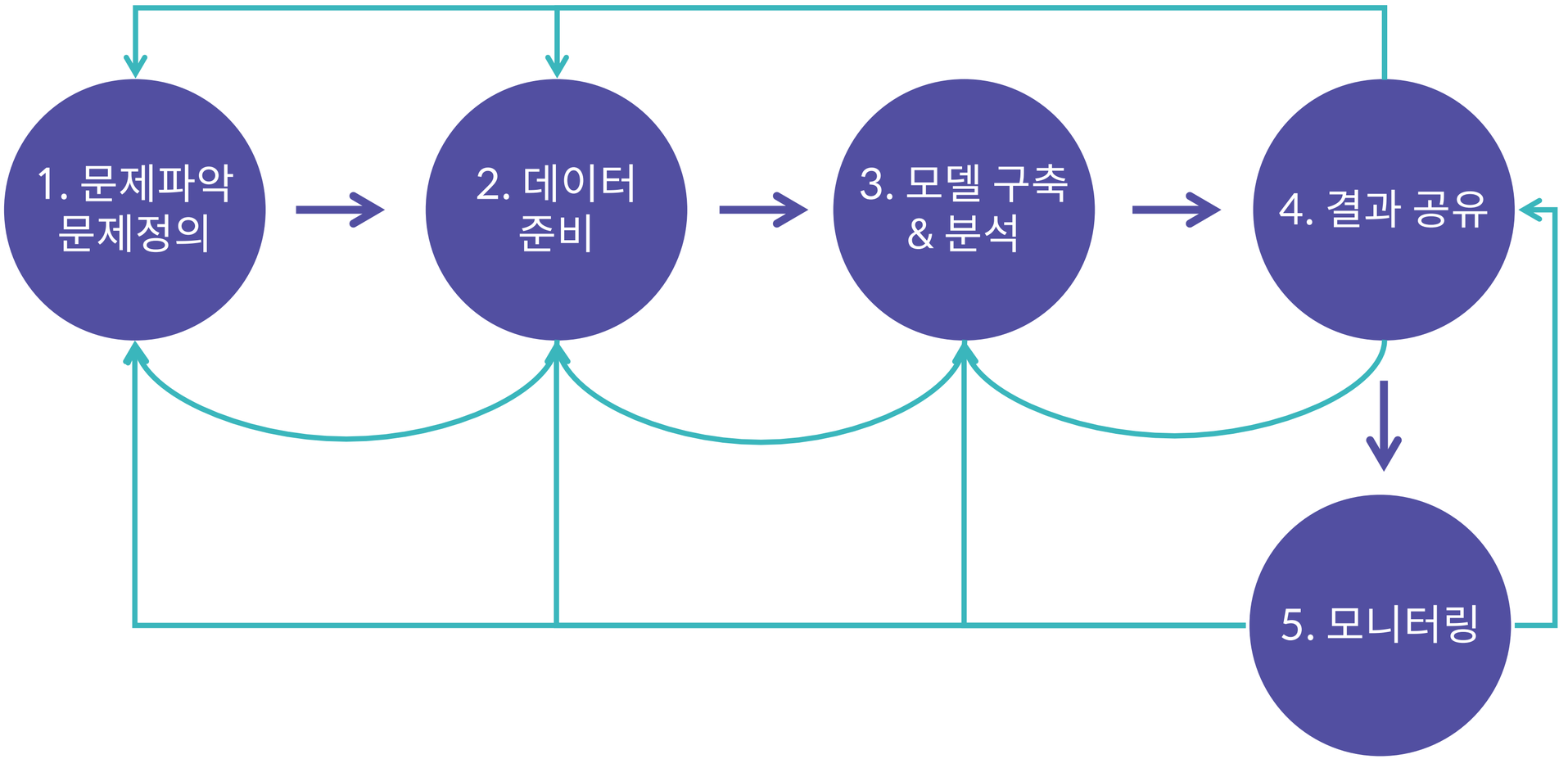
3. 데이터 과학자의 업무 살펴보기

데이터 과학은 예술이다?

이상적인 머신러닝 업무 프로세스



현실의 머신러닝 업무 프로세스



1. 문제파악 및 문제정의

1) **비즈니스** 문제 파악
2) **머신러닝** 문제로 전환
3) 머신러닝 도입 **필요성/가능성** 체크
4) 도입에 따른 **효과검증** 설계

2. 데이터 준비

1) 가능한 다양하고 **많은** 데이터 확보
2) 머신러닝을 도입할 **시스템 설계**
3) 데이터 분석 및 이해 – **Understanding**
4) 데이터 분석 및 이해 – **Preprocessing**
5) 데이터 분석 및 이해 – **Exploring**
6) **Feature Engineering**
7) **학습, 검증, 테스트** 데이터셋 생성

3. 머신러닝 모델 구축 & 분석

1) 사용할 **모델/알고리즘** 선택
2) 실무적 **제약사항** 고려
3) **하이퍼파라미터** 설정
4) 모델 **학습**

4. 결과 공유

A. 코드 **배포** (Productionize)

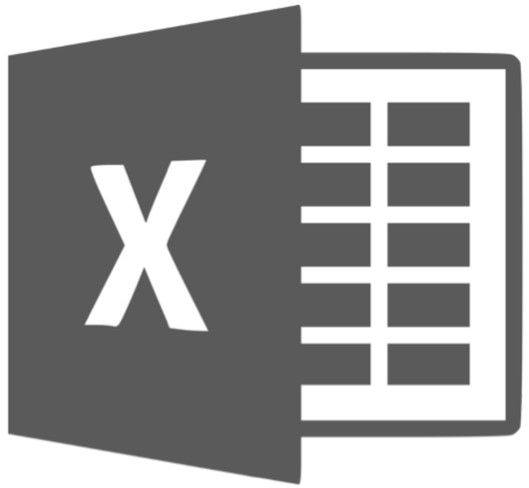
B. 보고서 작성, **결과정리** 및 발표

5. 모니터링

1) 모델의 **성능**을 지속적으로 tracking
2) **효과검증** 결과 tracking
3) 지속적인 **유지·보수** 계획/실행

머신러닝을 위한 데이터 과학자의 도구

Excel



머신러닝을 위한 데이터 과학자의 도구

Python

R

Python 머신러닝 Tool Box

IPython

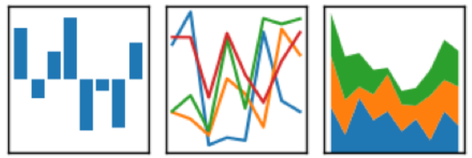
jupyter

NumPy

SciPy

pandas

$y_{it} = \beta' x_{it} + \mu_i + \epsilon_{it}$



matplotlib

scikit learn

TensorFlow