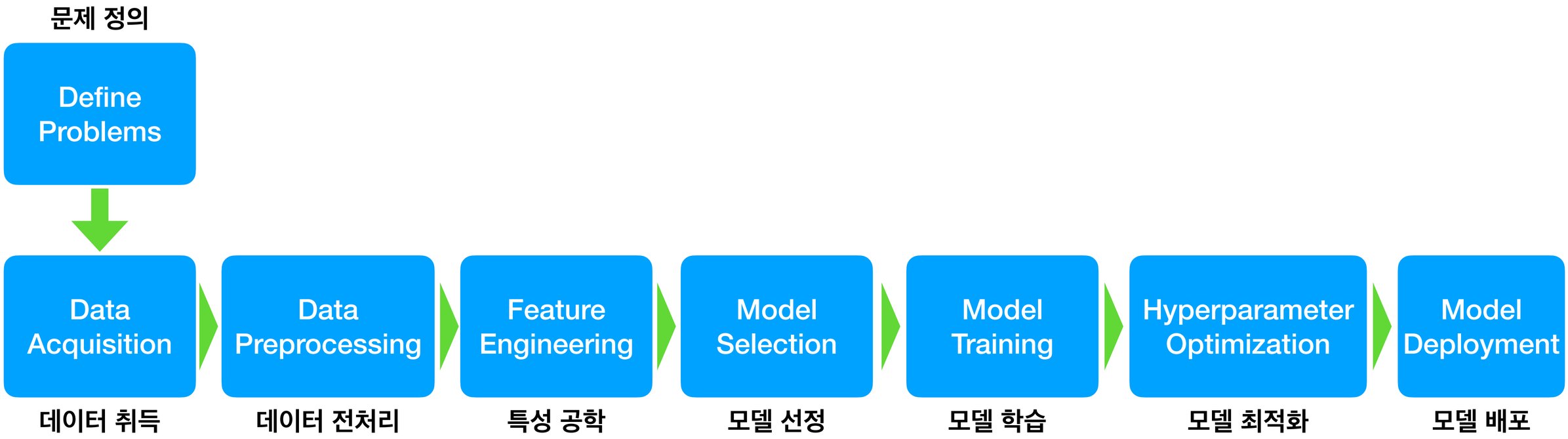
Chapter 4. ML pipeline

머신 러닝 모델 개발 프로세스

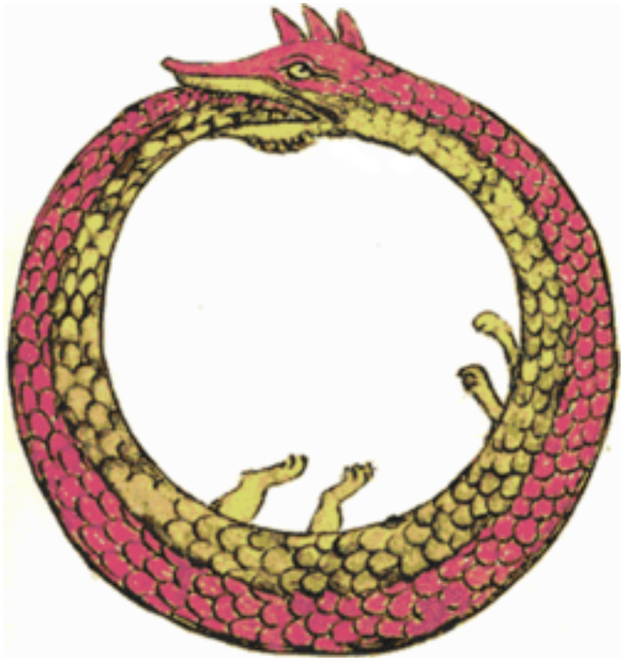


문제 정의 및 모델링

- 어떤 문제를 해결하고 싶은지?
  - 이미지를 통해 불량품을 탐지하고 싶다.
  - 구매 이력을 바탕으로 상품을 추천해주고 싶다.
- 풀고자 하는 문제의 유형은?
  - 대부분 Regression, Classification, Clustering
  - Regression : 연속적인 값을 예측하는 문제 (e.g., 주택 가격, 기온 예측 등)
  - Classification : 주어진 클래스 중 하나로 예측하는 문제 (e.g., 사물 분류 등)
  - Clustering : 데이터를 몇 개의 그룹 또는 클러스터로 나누는 문제 (e.g., 고객 그룹)

데이터 획득

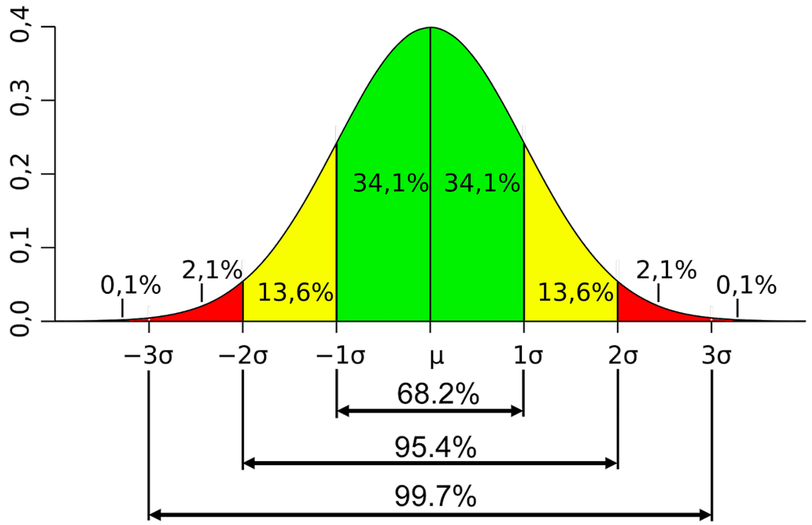
- 데이터 획득
  - 문제정의 과정에서 대략적인 데이터의 유형을 파악 가능
  - 데이터 수집 과정 및 저장소에서 데이터 가져오기
  - 데이터의 형태, 구조 및 특성 파악
- 데이터가 먼저냐? 문제가 먼저냐?



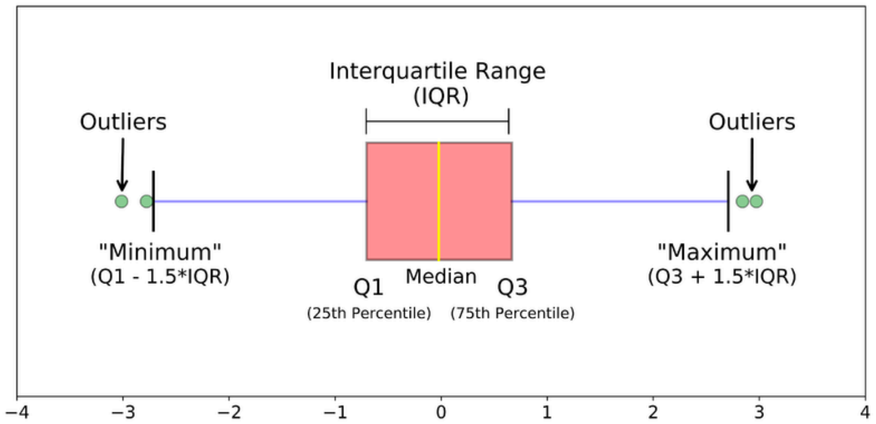
우로보로스

데이터 전처리 (결측치 및 이상치)

- 결측치 처리: 결측치 확인 및 대체 또는 제거
  - Interpolation, 평균, 중앙값, 최빈값 등 치환, 결측값 구분
- 이상치 처리: 이상치 탐지 및 대체 또는 제거



Z-score



Box plot

데이터 전처리 (정규화 및 인코딩)

- 데이터 정규화: 데이터 범위 변환
  - Z-score normalization : 데이터의 평균과 표준편차를 각각 0,1이 되도록 보정
  - min-max normalization : 데이터의 최소값과 최대값이 각각 0, 1이 되도록 보정
  - Robust scaler, power transformer, 로그 변환 등
- 범주형 데이터 인코딩:
  - 레이블 인코딩 : 범주형 레이블에 대한 숫자 레이블 할당 (e.g., ‘개’:1, ‘고양이’ :2)
  - 원 핫(One-hot) 인코딩 : 전체 클래스에 대한 벡터를 생성 후 해당 클래스에 1 할당.
    (e.g., ‘개’, ‘고양이’, ‘토끼’. → ‘개’ = [1,0,0], ‘고양이’ = [0,1,0], ‘토끼’ = [0,0,1].

특성 공학 (Feature engineering)

- 다항 변환 (Polynomial feature) : 기존 변수들의 다항식 조합 활용, (e.g., x^2, xy, y^3).
- 그룹 특성 (Group feature) : 연관된 변수 그룹이 존재 시 통계값 활용, (e.g., $t_1, t_2, t_3, \cdots, t_k \rightarrow \mu_t, \sigma_t$)
- 구간화 (Binning) : 연속형 변수를 구간을 나누어 범주형 변수로 변환
  - (e.g., 0~200까지 속도값, → stop, slow, fast, very fast 클래스로 구분)
- 변수 분할 (Feature split) : 복합적으로 구성된 변수를 분할하여 활용
  - (e.g., 20220301 → (연도) 2022, (월) 3, (일) 1.
- 푸리에 및 웨이브렛 변환 (시계열 데이터), 변수 추가 (e.g., 날짜데이터 → 공휴일 변수), PCA 등

모델 선정

- 주어진 문제 구성과 데이터에 적합한 기계학습 모델 선정
  - K-means 클러스터링 → 시계열 예측 ?
  - 시계열 데이터 → 그래프 이미지 변환 → 2D CNN ?
  - 가능하지만 정석적인 접근법부터 시작.
- 딥러닝 이후에는…
  - 이미지, 자연어, 시계열 데이터 → 딥러닝 모델들
  - 표 형식의 데이터 → LightGBM, Catboost 등 Gradient boosting 모델들

모델 학습

- 모델 선정 이후에는 모델의 학습이 필요
- 반복적인 학습 과정을 통해 점진적인 성능의 향상
- 학습 과정과 개념은 기계학습 알고리즘 별로 차이가 있음

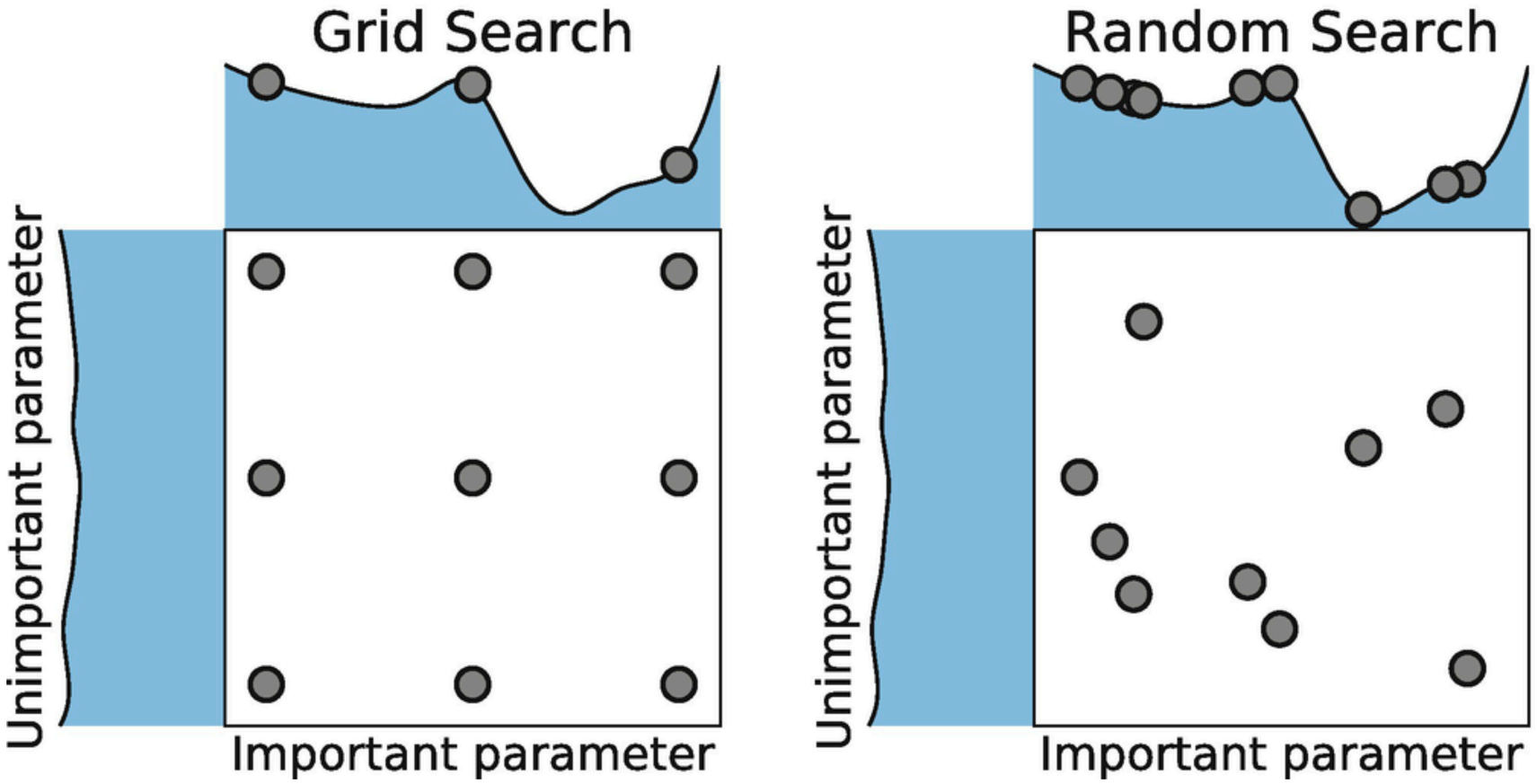
하이퍼 파라미터 튜닝

- 하이퍼 파라미터 :
  - 학습하고자 하는 머신러닝 모델에 대해서 모델 구조, 학습 과정 등과 관련하여 사전에 설정해주는 변수들.
  - 학습을 통해 배우는 파라미터는 아니지만, 모델 학습 결과에 큰 영향을 미침.
- 머신러닝 모델 별 하이퍼 파라미터 예시
  - 신경망 : 뉴런의 수, 레이어의 수, 모델 구조, 활성 함수, learning rate, optimizer, etc.
  - 서포트벡터머신 : 커널의 종류, 마진(margin)
  - 의사결정트리, 랜덤포레스트, lightGBM : 트리의 수, 최대 깊이, 서브샘플링 여부 등

하이퍼 파라미터 튜닝

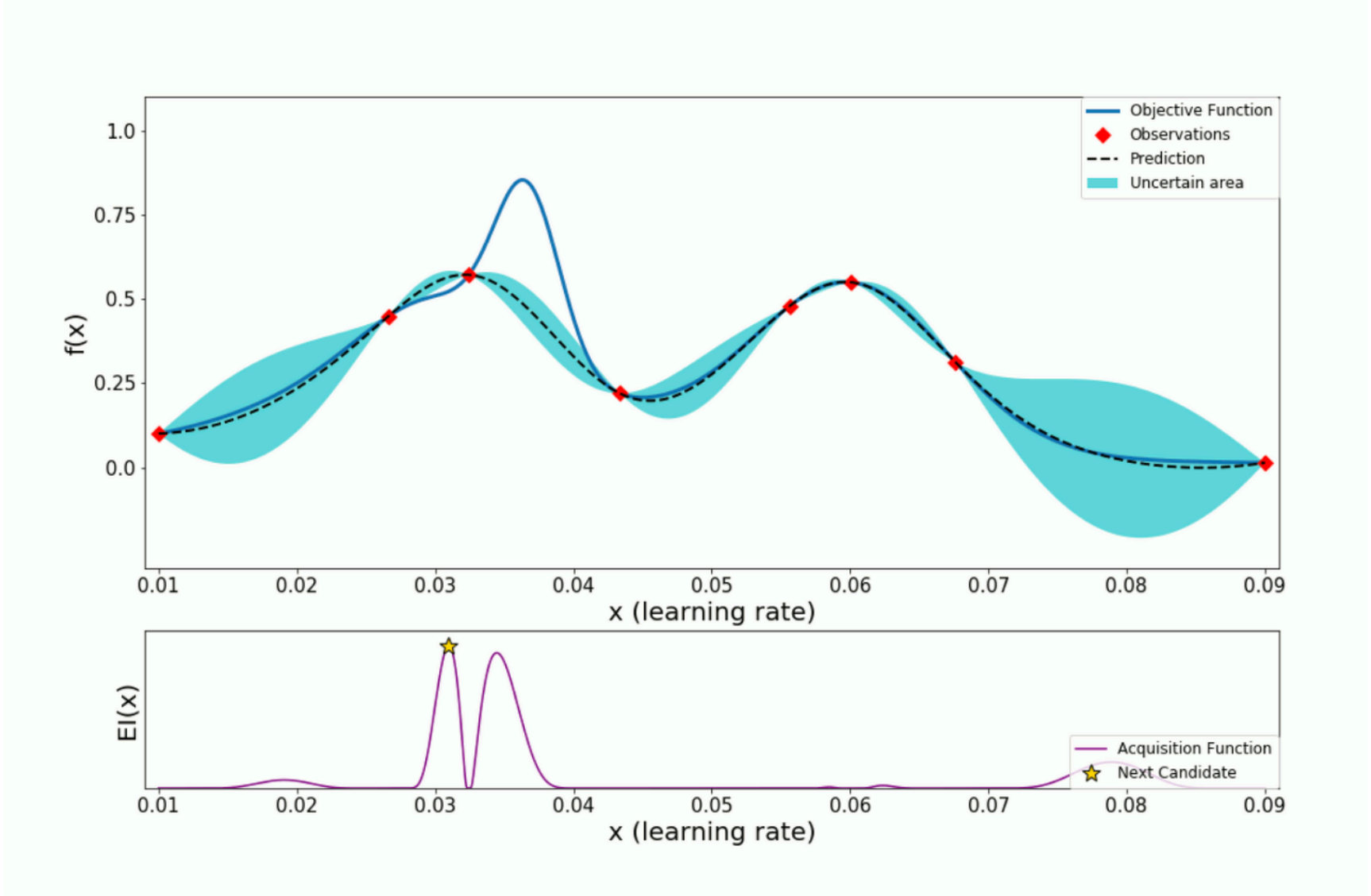
- 최적의 하이퍼 파라미터를 선정하기 위해서는?
  - 모의고사가 필요함, 즉 밸리데이션 셋에 대한 성능을 바탕으로 하이퍼 파라미터 탐색.
  - 테스트 셋으로 하이퍼 파라미터 선정시 테스트셋에 과적합 될 수 있음.
- 대표적인 방법론
  - Manual search, Grid search, Random search, Bayesian optimization

Grid search & Random search



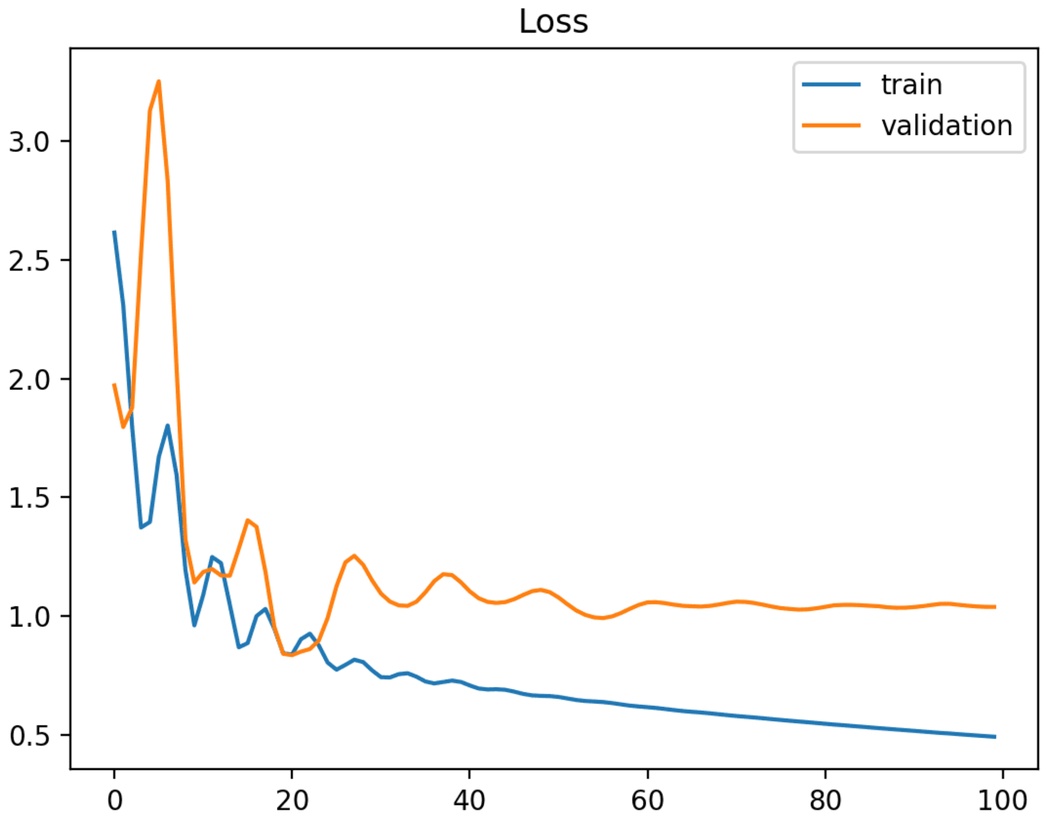
Bergstra, James, and Yoshua Bengio. "Random search for hyper-parameter optimization." The Journal of Machine Learning Research 13.1 (2012): 281-305.

Bayesian optimization



성능 평가

- 주어진 태스크에 대한 모델의 성능은 어떻게 판단할 수 있을까?
- MLE >> loss를 통한 평가?
- 일반적으로 loss는 모델 학습 과정에서 학습이 잘 진행되는지 살펴보는데 사용



태스크에 따른 평가지표

- 태스크에 따라 다양한 평가지표를 활용
- Classification : Accuracy, F1 score, AUROC, PRAUC 등
  - 분류 지표의 경우, 분류 기준(threshold)에 따라 지표의 변동 가능
- Regression : Mean absolute error, Mean squared error, R-squared 등
  - 오차에 기반하므로 보통 작을수록 좋음.

모델 개발의 목적

- 이러한 평가 지표는 무엇을 대상으로 계산하여야할까?
- 최종적인 목적은 학습된 모델을 실제 서비스에 적용하는 것에 있음.
- 즉 모델에 대한 평가는 실제 서비스에 적용했을 때의 성능에 대한 것임
- 실제 모델 배포 전에 개발/학습한 모델이 좋은 모델인지 어떻게 평가 할 수 있을까?

모델의 일반화 성능

- 온실 속에서 키운 모델을 온실 밖으로 보냈을때 잘 할 수 있을까?
- 모델의 일반화 성능(generalization performance)이란?
  - 이전에 보지 못한 새로운 데이터를 적용시 모델이 갖는 성능
  - a) 실제 서비스에 적용 후 데이터 분석 → 가능하지만 단점이 많음.
  - b) 실제 서비스 적용 전에 일반화 성능을 유추 할 수 있는 방법?

대학 입시 과정을 통한 예시

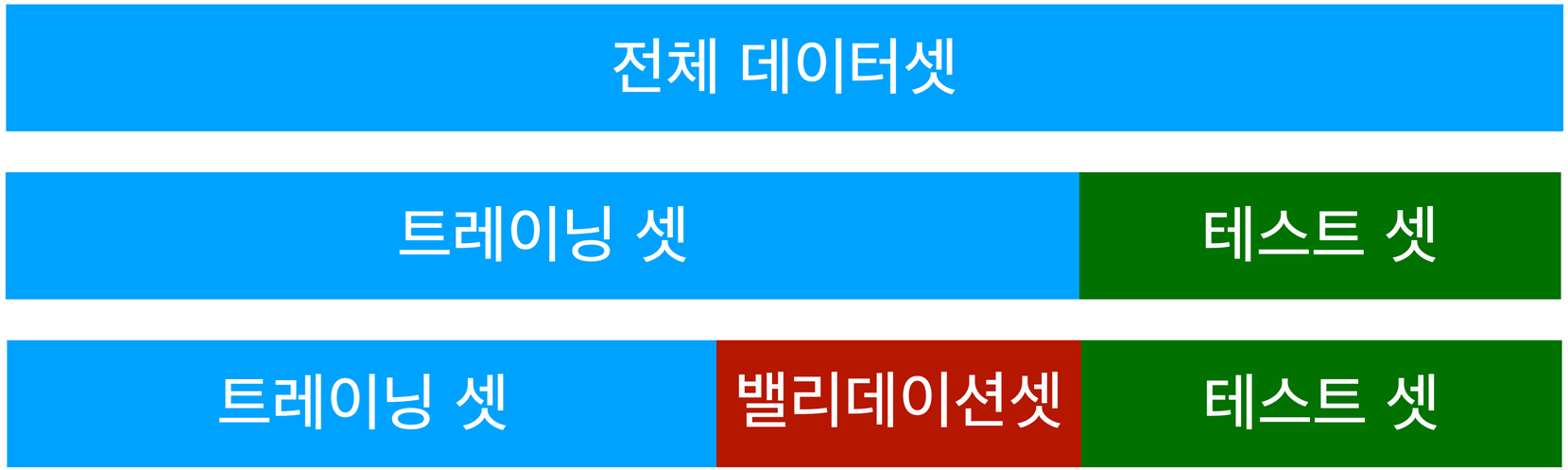
구성 요소들

- 수학능력평가시험 : 대학에서의 교육과정을 얼마나 잘 수학(修學)할 수 있는지 평가하는 것이 시험의 목적
- 수능 공부 : 수학능력평가 시험을 잘 보기위한 학습
- 모의 고사 : 수능 시험을 잘 볼 수 있도록 학습이 잘 진행되고 있는지 체크

대학 입시 과정과 기계학습 모델 학습 과정

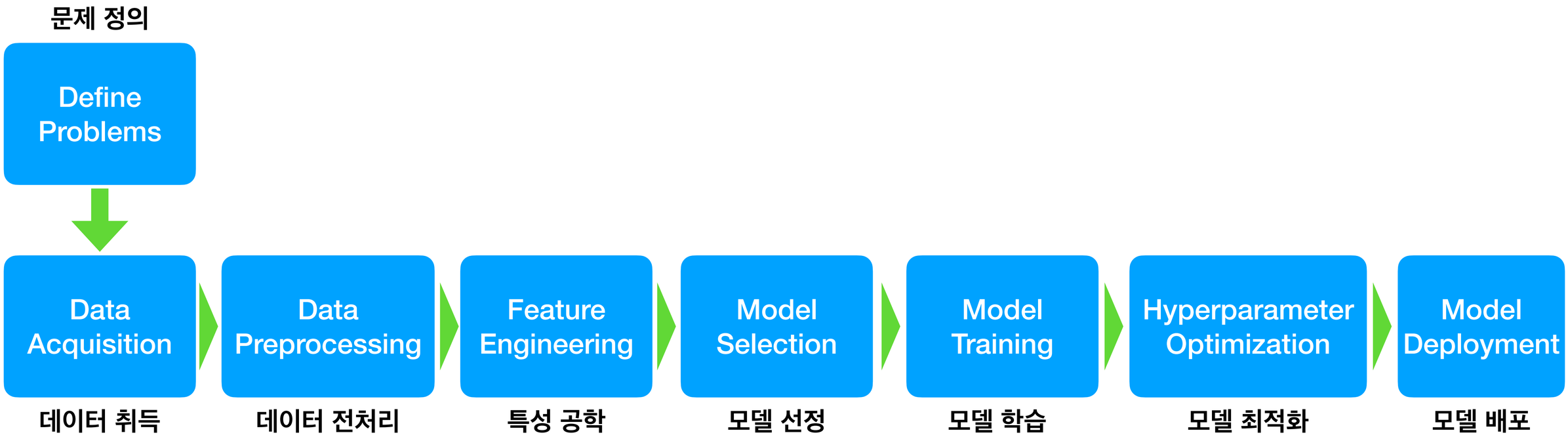
- 사람 : 모델
- 대학에서의 공부 : 실제 서비스 적용 시의 문제
- 수능 : 테스트 셋 (test set)
- 모의고사 : 밸리데이션 셋 (validation set)
- 공부 : 트레이닝 셋 (training set)
- 수능문제가 유출된다면? → Data leakage를 주의해야 함.

Hold-out validation

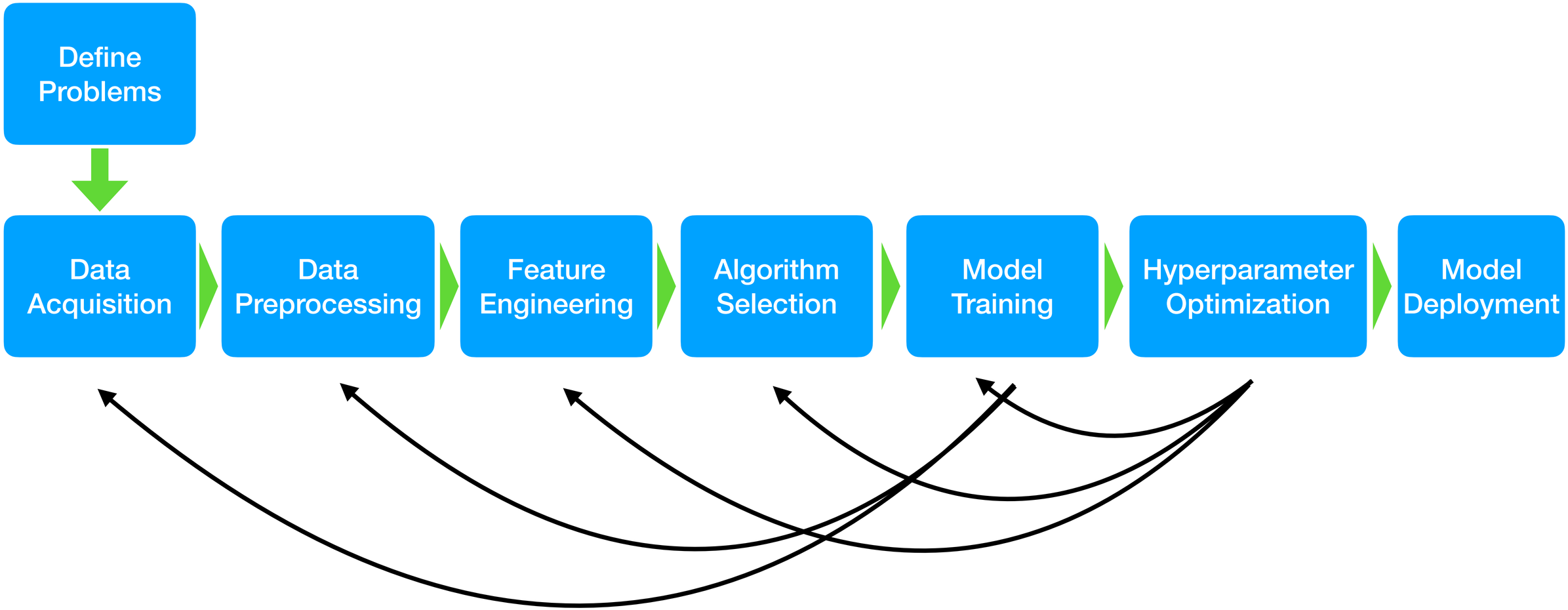


7:3, 6:2:2

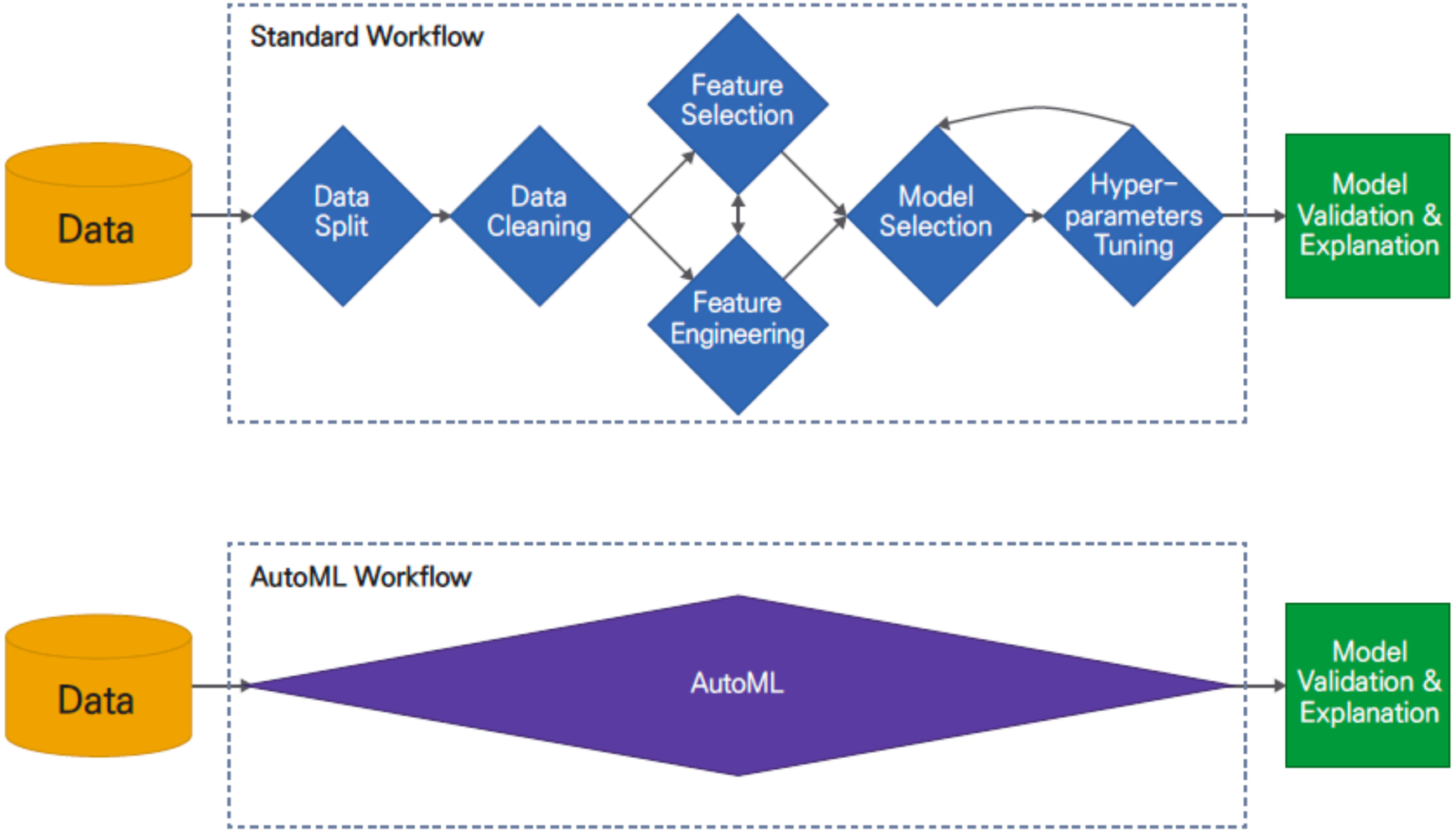
머신 러닝 모델 개발 프로세스



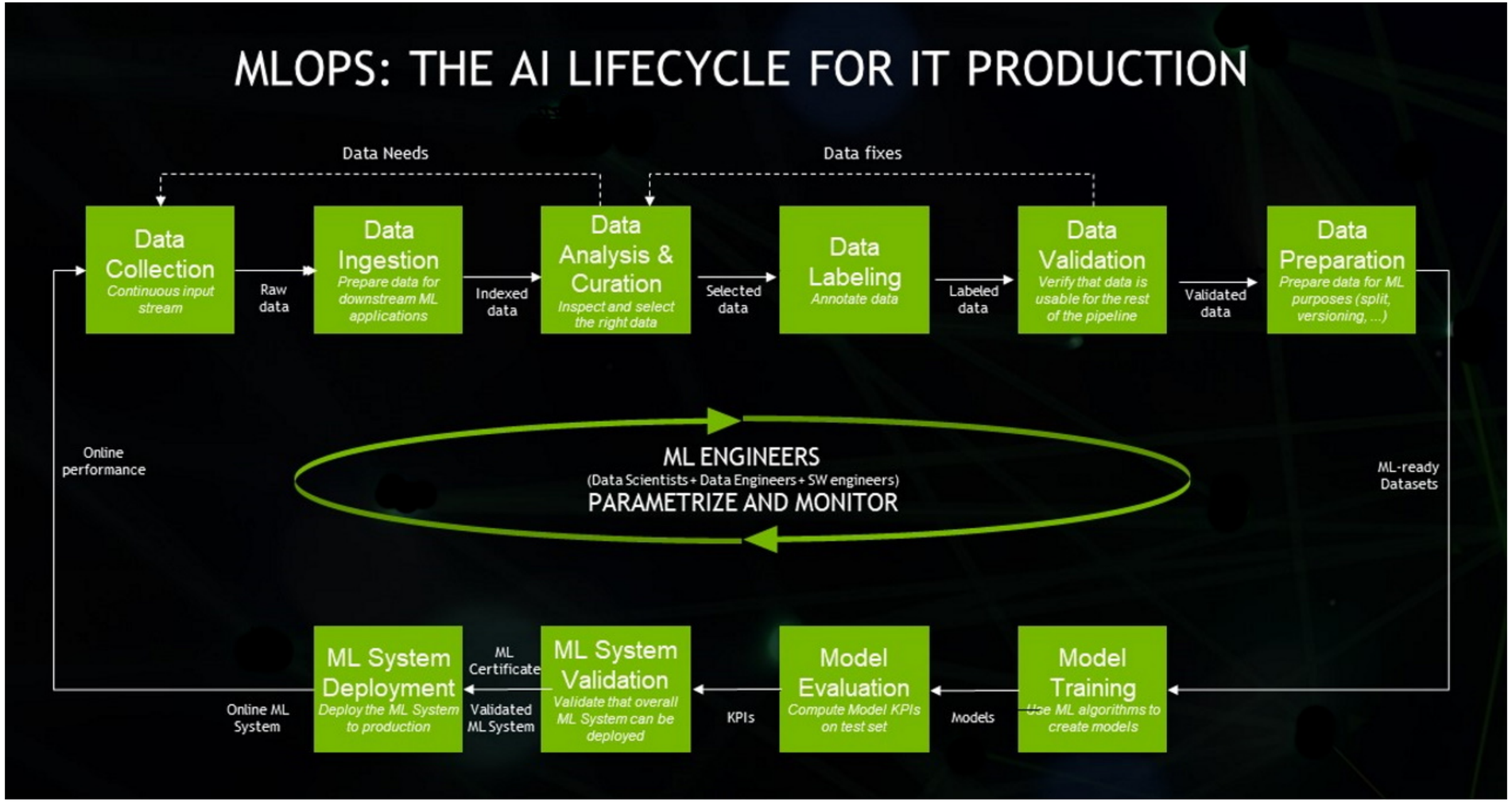
단방향이 아닌 과정...


Define Problems
Data Acquisition
Data Preprocessing
Feature Engineering
Algorithm Selection
Model Training
Hyperparameter Optimization
Model Deployment

Automation of Machine Learning



Auto ML & MLOps



https://blogs.nvidia.com/blog/what-is-mlops/

Auto ML frameworks

mljar

Machine Learning for Humans

Google cloud AutoML

PYCARET

scikit learn

auto-sklearn

H_2O AutoML

AutoGluon

AutoKeras