

Let's go MLOps!



Be a MLOps Specialist!

C1_W2: Introduction to Machine Learning in Production

Select and Train a Model

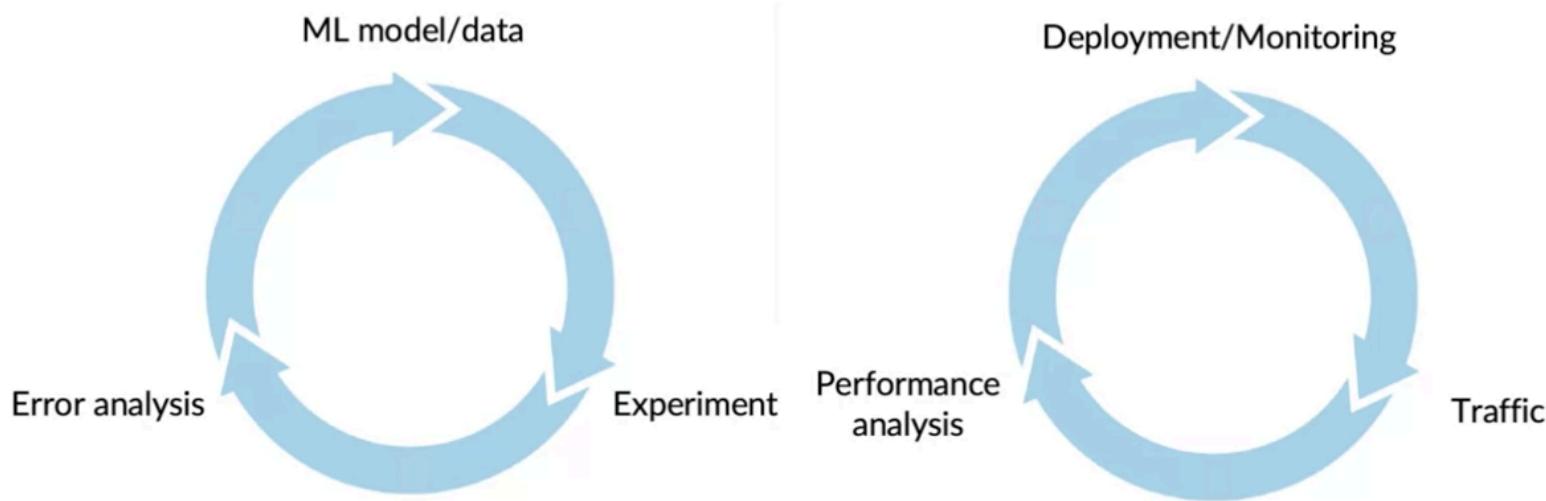
MLOps 이영전



본 PPT 슬라이드 제작에 사용된 PPT 자료들은 Coursera MLOps Specialist Course1강의를 참고했습니다.

- Introduction to Machine Learning in Production:
<https://www.coursera.org/learn/introduction-to-machine-learning-in-production>

Just as ML modeling is iterative, so is deployment



Iterative process to choose the right set of metrics to monitor.

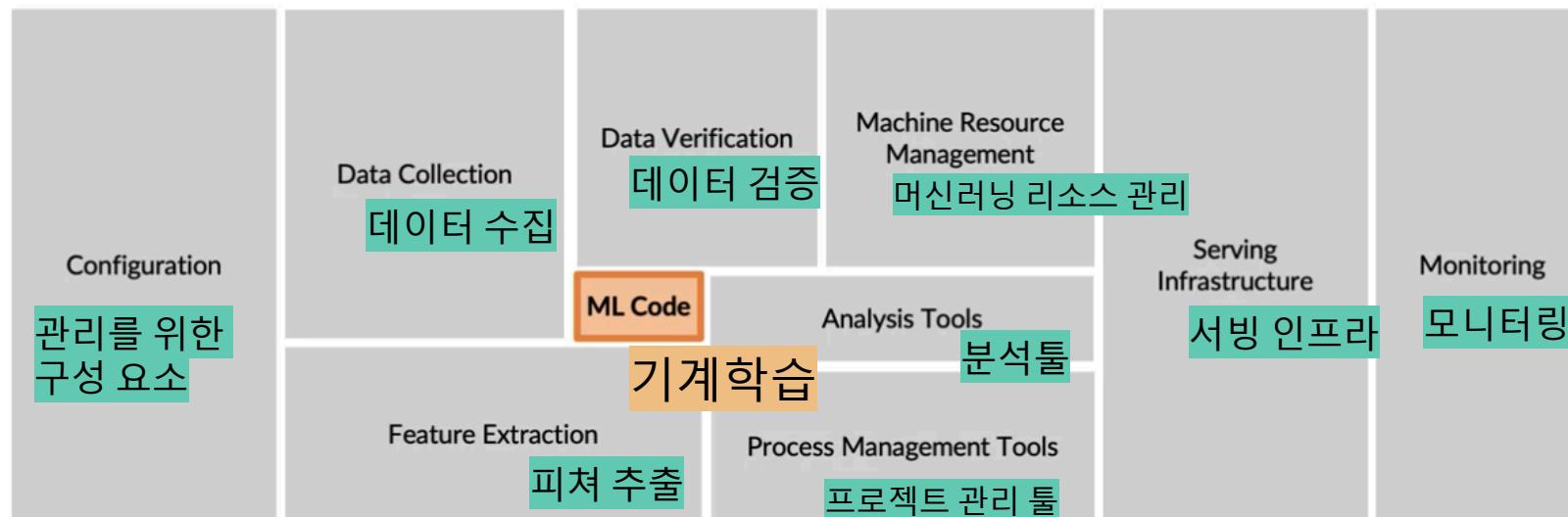


머신러닝 모델링과 같이 머신러닝 프로덕트 개발도 반복적인 프로세스로 구성되어 있습니다

ML in production

머신러닝 프로젝트 관점에서 바라봤을 때, 머신러닝 코드(Machine Learning Code)는 전체 프로젝트 코드의 일부분에 불과합니다

The requirements surrounding ML infrastructure

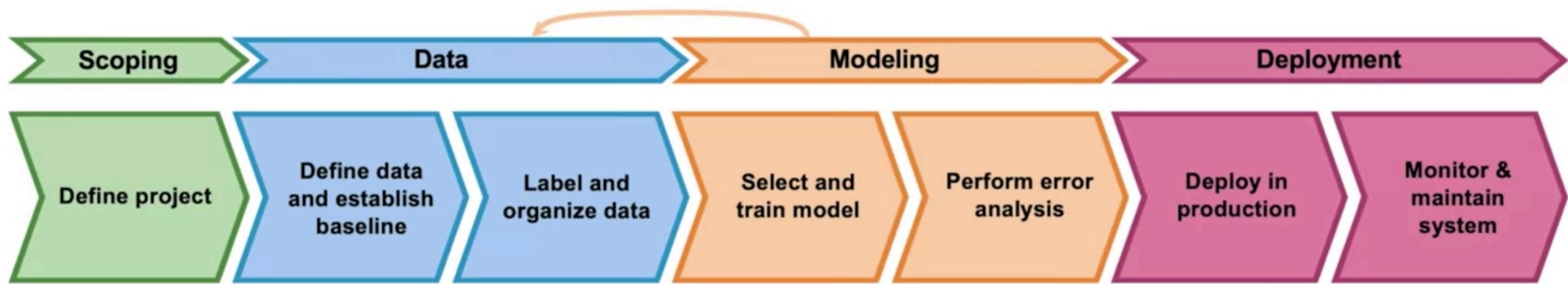


[D. Sculley et. al. NIPS 2015: Hidden Technical Debt in Machine Learning Systems]

The ML project lifecycle

The ML project lifecycle

머신러닝 프로젝트 라이프사이클은 다음과 같습니다



Scoping:
문제를 정의하는 단계

- Data:**
- 데이터에 대한 정의와 베이스 라인 설계
 - 데이터 레이블링 및 정리 (organize)

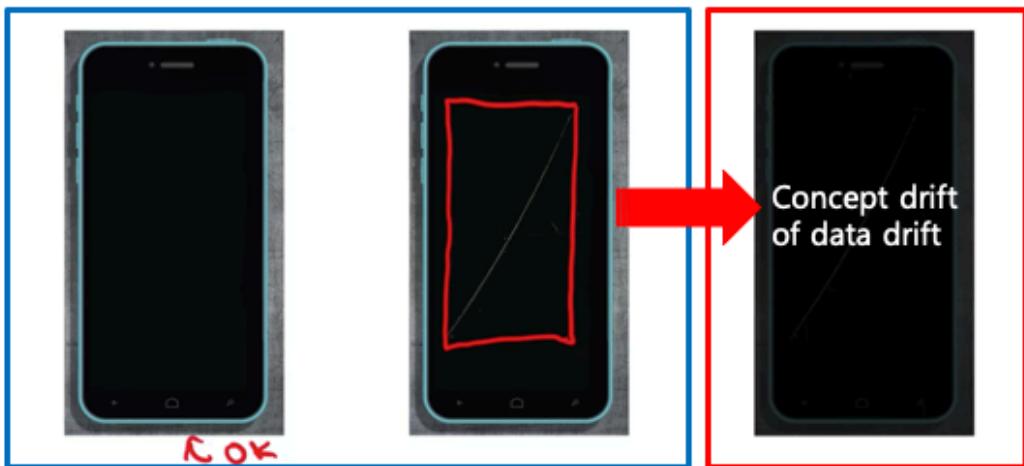
- Modeling**
- 모델 선택 및 학습
 - Error Analysis를 통한 모델 성능 검증

- Deployment**
- 배포 단계
 - 프로덕트 배포
 - 모니터링과 시스템 관리 (e.g. 데이터 분포가 바뀌었을 경우, 모델 업데이트)

Data Drift, Concept Drift

Q1. 학습때 보지 못했던 데이터를 마주치게 된다면 어떤 일이 일어날까?

Visual inspection example



환한 조명에서 찍은 사진으로 모델을 학습했는데... → 어두운 조명에서 찍힌 사진이 입력값으로 들어왔다

A2. 이러한 문제를 concept drift 또는 data drift라고 한다

- **Concept drift** happens when the statistical properties of the target variable itself change. The meaning of what you are trying to predict changes and therefore the model will not work well for this updated definition. For example, the definition of what is considered a fraudulent transaction could change over time.

- **Data drift** happens when the statistical properties of the underlying variables that predict an outcome change. A classic example is the natural drift in data due to seasonality.

- Upstream Data Change happens when there is a change in the data pipeline upstream which has an impact on the model performance. For example, camera's being replaced and as a consequence the units of measurement change.

쉽게 생각하면...

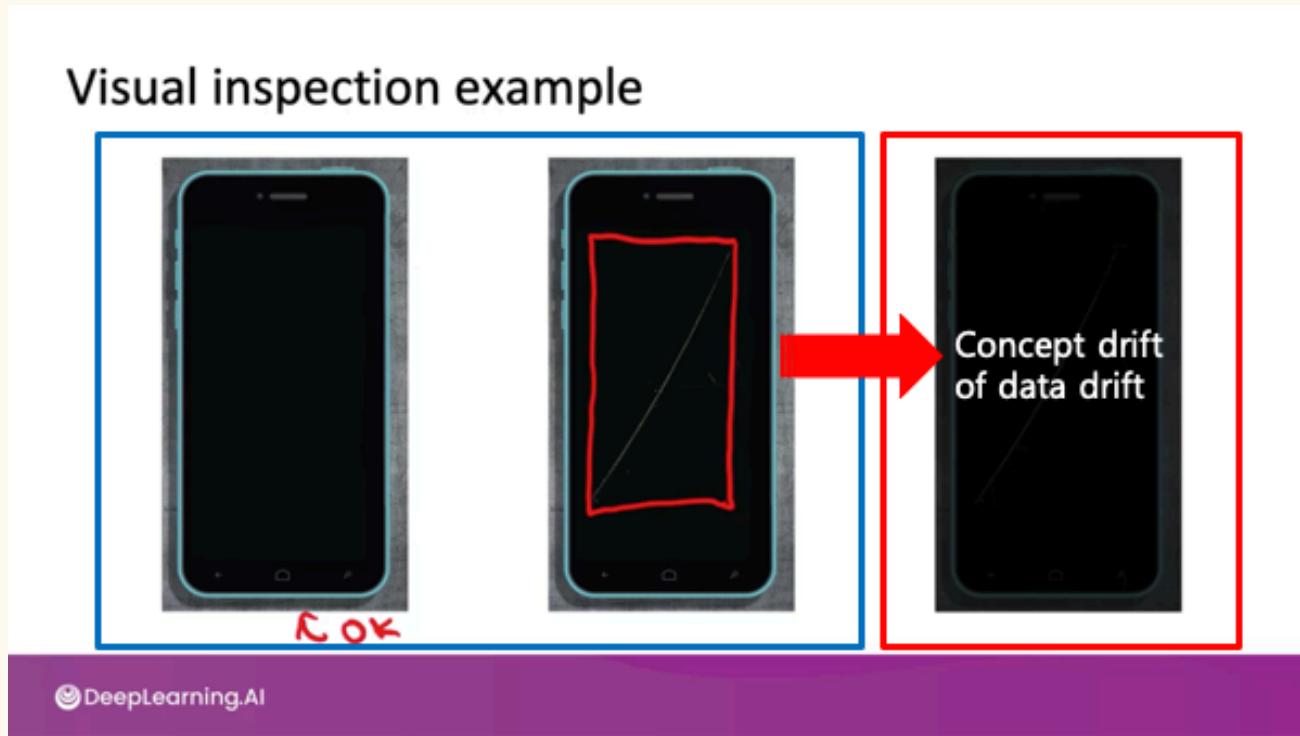
- concept drift: distribution of predictions
- data drift: data-and feature contribution distribution

- 데이터가 바뀌었네? ⇒ 데이터 드리프트
- 컨셉이 바뀌었네? ⇒ 컨셉 드리프트

<https://www.viqtordavis.com/en-us/media/data-drift#:~:text=Data%20drift%20is%20a%20change,current%20real%2Dtime%20production%20data>

concept drift or data drift

Q1. 학습때 보지 못했던 데이터를 마주치게 된다면 어떤 일이 일어날까요?



A2. 이러한 문제를 concept drift 또는 data drift라고 합니다

TL;DR

이번 시간에는 모델 개발과 관련된 모델 전략(model strategy)과 주요 과제(key challenge)를 다룹니다. 다양한 유형의 데이터를 작업하기 위한 오류 분석(error analysis) 및 전략(model strategy)을 다룹니다. 또한 클래스 불균형(class imbalance)과 심하게 왜곡된 데이터 세트(highly skewed data set)에 대처하는 방법을 다룹니다.

- # Model Strategy
- # Key Challenge
- # Error Analysis
- # Class Imbalance problems
- # Highly Skewed Dataset

Goal

- Identify the key challenges in model development.
- Describe how performance on a small set of disproportionately important examples may be more crucial than performance on the majority of examples.
- Explain how rare classes in your training data can affect performance.
- Define three ways of establishing a baseline for your performance.
- Define structured vs. unstructured data.
- Identify when to consider deployment constraints when choosing a model.



Goal

- List the steps involved in getting started with ML modeling.
- Describe the iterative process for error analysis.
- Identify the key factors in deciding what to prioritize when working to improve model accuracy.
- Describe methods you might use for data augmentation given audio data vs. image data.
- Explain the problems you can have training on a highly skewed dataset.
- Identify a use case in which adding more data to your training dataset could actually hurt performance.
- Describe the key components of experiment tracking.



학습 목표

- 모델 개발의 주요 과제를 학습합니다.
- 불균형적으로 중요한 소수의 예에 대한 성능이 (performance on small set of disproportionately important examples) 대다수의 예에 대한 성능 (performance on the majority of examples)보다 더 중요할 수 있는 방법을 설명합니다.
- 훈련 데이터의 드문 클래스(rare class)가 성능에 어떤 영향을 미칠 수 있는지 설명합니다.
- 모델 퍼포먼스에 대한 기준(baseline for your performance)을 설정하는 세 가지 방법을 정의합니다.
- 정형 데이터와 비정형 데이터를 정의합니다.
- 모델을 선택할 때 배포 제약 조건을 고려해야 하는 경우를 확인합니다.



학습 목표

- ML 모델링 시작과 관련된 단계를 나열합니다.
- 오류 분석을 위한 반복 프로세스를 설명합니다.
- 모델 정확도를 개선하기 위해 작업할 때 우선 순위를 정할 항목을 결정하는 핵심 요소를 식별합니다.
- 오디오 데이터와 이미지 데이터를 비교하여 데이터 증대에 사용할 수 있는 방법을 설명합니다.
- 심하게 치우친 데이터 세트에 대해 학습할 수 있는 문제를 설명합니다.
- 훈련 데이터 세트에 더 많은 데이터를 추가하면 실제로 성능이 저하될 수 있는 사용 사례를 식별합니다.
- 실험 추적의 주요 구성 요소를 설명합니다.



Overview

01 Selecting and Training a Model

02 Error analysis and performance auditing

03 Data iteration



1

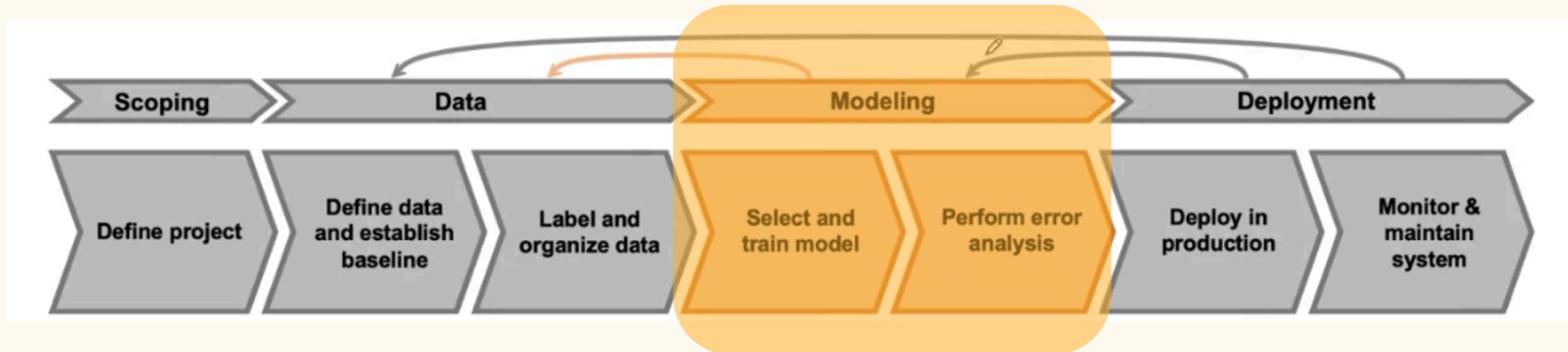
Selecting and Training a Model



- Key challenges
- Why low average error isn't good enough
- Establish a baseline
- Tips for getting started

Modeling Overview

이번 시간에는 기계학습 모델 배포에 적합한 머신러닝 모델 구축을 위한 모범 사례에 대해 학습해보겠습니다



some of the key challenges of trying to build a production-ready machine learning model

- e.g.
 - *how do you handle new datasets?*
 - *what if you do well in the test set, but for some reason, that still isn't good enough for your actual application?*

Key challenges

- AI System
- Model development is an iterative process
- Challenges in model development

**AI system = Code + Data
(algorithm/model)**

“AI systems or machine learning systems comprise both code, meaning the algorithm or the model as well as data”

- Machine Learning System의 AI System은 코드와 데이터를 포함하는 것입니다
- 지난 수십년 동안 머신러닝 연구의 중심은 코드 개선을 통한 문제해결이었습니다
- 반면 Data-Centric 방식은 Data에 조금 더 많은 시간을 할애하고 집중하는 방법입니다

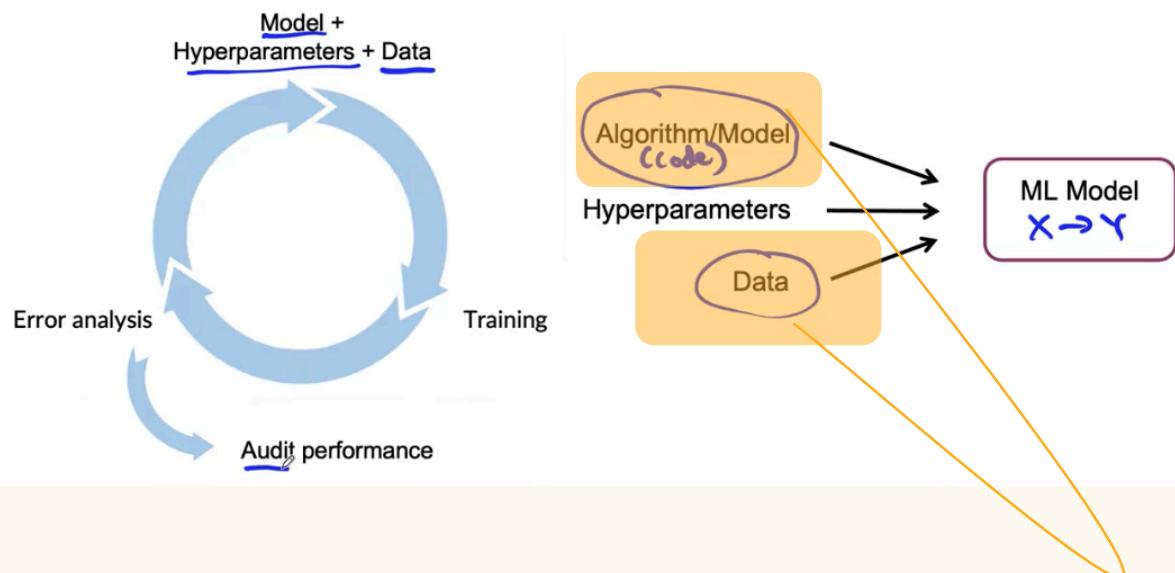
Key challenges

- AI System
- Model development is an iterative process
- Challenges in model development

Q. 모델 개발 프로세스는 어떻게 될까?

A. 모델 개발은 매우 반복적인 프로세스입니다

Model development is an iterative process



- 일반적으로 모델 개발 과정은 왼쪽 그림과 같습니다
- 기계학습은 이러한 경험적 프로세스이기 때문에 학습 루프를 매우 빠르게 여러번 통과하는 것이 성능 향상의 핵심입니다
- 성능 향상을 위한 방법으로는 다음의 방법들이 있습니다
 - 매번 데이터를 수정하는 방법
 - 모델을 수정하는 방법
 - 하이퍼파라미터를 수정하는 방법
- 위의 과정을 충분히 수행하여 좋은 모델을 얻은 후 더 많은 오류 분석과 최종 감사(final audit)을 거쳐 작동을 확인합니다

알고리즘(Code)/모델과
데이터에 집중해보겠습니다!

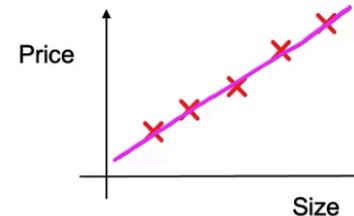
Key challenges

- AI System
- Model development is an iterative process
- Challenges in model development

“AI 시스템 개발의 어려운 점들에 대해 알아본다”

Challenges in model development

1. Doing well on training set (usually measured by average training error).



2. Doing well on dev/test sets.

3. Doing well on business metrics/project goals.

AI 시스템 개발의 어려운 점으로 어떤 점들이 있을까요?
모델 개발에 있어서의 주요 challenge는 다음과 같습니다

1. training set에 대해 성능이 좋아야 한다
2. dev/test set에 대해 성능이 좋아야 한다
3. business goal 또는 project goal을 달성해야 한다

실제 대부분의 경우 dev/test set에 대해 좋은 성능을 달성하더라도 business metric과 project goal을 달성하지 못하는 경우가 많다고 합니다.

Key challenges

- AI System
- Model development is an iterative process
- Challenges in model development

dev/test set에 대한 성능이 좋더라도...

Unfortunate conversation in many companies



MLE: "I did well on the test set!"

← test set으로 실험했는데 완전 좋았어요!



Product Owner: "But this doesn't work for my application"

← 그런데 잘 작동되지 않는데요?



MLE: "But... I did well on the test set!"

← 아니에요... test set에서 잘 됐다고요...!



Key challenges

- AI System
- Model development is an iterative process
- Challenges in model development

*AI System 개발 프로젝트 진행에 있어서 low average test error 가 충분한 지표가 아닐 경우
어떻게 하면 좋을지, 다음 챕터에서 다뤄보도록 하겠습니다!*

Why low average error isn't good enough

지금까지 우리는 Machine Learning System이란 무엇인지 알아봤습니다.
특히 머신러닝 모델 배포는 모델 개발 단계와 마찬가지로 반복적인 과정으로 이루어져 있다는 것을 배웠습니다.

하지만 **문제**가 발생했습니다. 머신러닝 개발팀에서 low dev/test error를 달성했더라도 실제 프로덕션 환경에서는 적용이 되지 않을 수 있다는 것입니다.

이러한 경우 어떻게 해야 할지 알아보도록 하겠습니다.

Why low average error isn't good enough

- Performance on disproportionately important examples
- Performance on key slices of the dataset
- Rare classes
- Unfortunate conversation in many companies

TL;DR

머신러닝 시스템이 test dataset에 대해 낮은 오류를 갖더라도 불균형적으로 중요한 일련의 예제에 대한 성능이 충분하지 않으면 기계 학습 시스템은 여전히 프로덕션 배포에 적합하지 않습니다.

Performance on disproportionately important examples



Web Search example

"Apple pie recipe" "Latest movies"
 "Wireless data plan" "Diwali festival"

} Informational and Transactional queries

"Stanford" "Reddit" "Youtube"

} Navigational queries

• **Informational and Transactional queries:** 웹 검색 엔진은 가장 관련성이 높은 결과를 반환하고 싶지만 사용자는 기꺼이 최상의 결과에 대해서 순위를 매길 수 있습니다. e.g. 1순위, 2순위, 3순위...

• **Navigational queries:** 의도가 명확한 쿼리입니다. Navigational query에서는 의도에서 벗어난 결과에 대해 보다 덜 관용적입니다

➔ Navigational queries are Disproportionately Important set of example

Why low average error isn't good enough

- Performance on disproportionately important examples
- Performance on key slices of the dataset
- Rare classes
- Unfortunate conversation in many companies

Performance on key slices of the dataset

Example: ML for loan approval

Make sure not to discriminate by ethnicity, gender, location, language or other protected attributes.

Example: Product recommendations from retailers

Be careful to treat fairly all major user, retailer, and product categories.

you will probably want to make sure that your system does not unfairly discriminate against loan applicants according to their ethnicity, gender, maybe their location, their language, or other protected attributes.

→ 여러분은 여러분의 AI System이 편향된 결과를 반환하기를 원하지 않을 것입니다

Even if a learning algorithm for loan approval achieves high average test set accuracy, it would not be acceptable for production deployment if it exhibits an unacceptable level of bias or discrimination.

→ Test set에 대해 성능이 좋다고 하더라도 받아들일 수 없는 편향이 존재한다면 결과를 받아들일 수 없을 것입니다

Why low average error isn't good enough

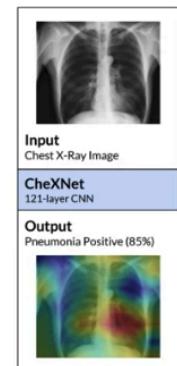
- Performance on disproportionately important examples
- Performance on key slices of the dataset
- Rare classes
- Unfortunate conversation in many companies

Rare classes

Skewed data distribution
99% negative 1% positive

Condition	Performance
Effusion	0.901
Edema	0.924
Mass	0.909
Hernia	0.851

→ 전체 데이터에서 100개에 불과한 Hernia를 모두 '음성'으로 판별하더라도 평균 테스트 정확도는 크게 손상되지 않을 수 있습니다



Q. *Hernia*와 같은 Rare classes의 결과를 어떻게 해석할 수 있을까요?
 Q. 만약 *Hernia*에 대해 모든 경우 '음성'으로 판별할 경우 결과는 어떻게 될까요?

만약 머신러닝 모델(또는 알고리즘)이 *Hernia*의 모든 사례를 완전히 무시할 수 있다면 이 경우 평균 테스트 정확도는 약간만 영향을 받을 수 있게 됩니다.

즉 *Hernia*가 얼마 발생하지 않기 때문에 대부분의 값을 '음성'으로 판별한다면 평균 테스트 정확도를 크게 손상시키지 않는 결과를 발생시킬 수 있는 것입니다.

이것이 바로 Rare class가 갖는 문제점입니다.

Why low average error isn't good enough

- Performance on disproportionately important examples
- Performance on key slices of the dataset
- Rare classes
- Unfortunate conversation in many companies



*Performance on disproportionately important examples,
Performance on key slices of the datasets,
Rare Classes*

...

Unfortunate conversation in many companies



MLE: "I did well on the test set!"



Product Owner: "But this doesn't work for my application"



MLE: "But... I did well on the test set!"

→ 결국 위의 조건에서 좋은 성능을 도출하지 못한다면 test set에 대해서 좋은 성능을 발휘하더라도 실제 프로덕트 환경에서는 좋은 성능을 발휘할 수 없게 될 수 있습니다

Establish a baseline

- Establishing a baseline level of performance
- Unstructured and structured data
- Ways to establish a baseline

Establishing a baseline level of performance

 **Speech recognition example:**

Type	Accuracy	Human level performance	HLP
Clear Speech	94%	→ 95%	10%
→ Car Noise	89%	→ 93%	4%
People Noise	87%	→ 89%	2%
→ Low Bandwidth	70%	→ 70%	~0%

 DeepLearning.AI

Q. 그렇다면 *baseline*은 어떻게 설정해야 하는 걸까?

HLP(Human lever performance)는 인간 수준의 성능을 의미합니다.

기계학습의 결과와 HLP를 비교하여 개선점을 찾을 수 있습니다. 어떤 Type의 데이터에 집중할 수 있는지 결정할 수 있는 것입니다

따라서 HLP를 *baseline*으로 여기는 것이 도움이 됩니다

예를 들어 Clear Speech 타입의 데이터에 대해 학습 모델의 정확도가 94%라고 가정합니다. 이 때, HLP의 정확도가 95%라면 1%의 성능 개선 여지가 있는 것입니다.

Establish a baseline

- Establishing a baseline level of performance
- Unstructured and structured data
- Ways to establish a baseline

Unstructured and structured data

Unstructured data	Structured data												
Image 	<table border="1"><thead><tr><th>User ID</th><th>Purchase</th><th>Number</th><th>Price</th></tr></thead><tbody><tr><td>3421</td><td>Blue shirt</td><td>5</td><td>\$20</td></tr><tr><td>612</td><td>Brown shoes</td><td>1</td><td>\$35</td></tr></tbody></table>	User ID	Purchase	Number	Price	3421	Blue shirt	5	\$20	612	Brown shoes	1	\$35
User ID	Purchase	Number	Price										
3421	Blue shirt	5	\$20										
612	Brown shoes	1	\$35										
Audio 													
Text This restaurant was great!	<table border="1"><thead><tr><th>Product ID</th><th>Product name</th><th>Inventory</th></tr></thead><tbody><tr><td>385</td><td>Football</td><td>158</td></tr><tr><td>477</td><td>Cricket bat</td><td>23</td></tr></tbody></table>	Product ID	Product name	Inventory	385	Football	158	477	Cricket bat	23			
Product ID	Product name	Inventory											
385	Football	158											
477	Cricket bat	23											

Unstructured data, Structured data에 따라 HLP를 베이스라인(baseline)으로 할지 여부를 정하는 것이 필요합니다.

일반적으로 Unstructured data의 경우 HLP를 baseline으로 하는 것이 도움이 됩니다.

Establish a baseline

- Establishing a baseline level of performance
- Unstructured and structured data
- Ways to establish a baseline

Ways to establish a baseline

- Human level performance (HLP)
- Literature search for state-of-the-art/open source
- Quick-and-dirty implementation
- Performance of older system

Baseline helps to indicates what might be possible. In some cases (such as HLP) is also gives a sense of what is irreducible error/Bayes error.

DeepLearning.AI

- HLP를 포함한 다양한 baseline 설정 방법들이 있습니다.
- 문헌 검색(Literature search)을 통한 최신/오픈소스 자료를 baseline으로 정할 수 있으며 빠른 구현, 이전 시스템 등을 baseline으로 정할 수 있습니다.

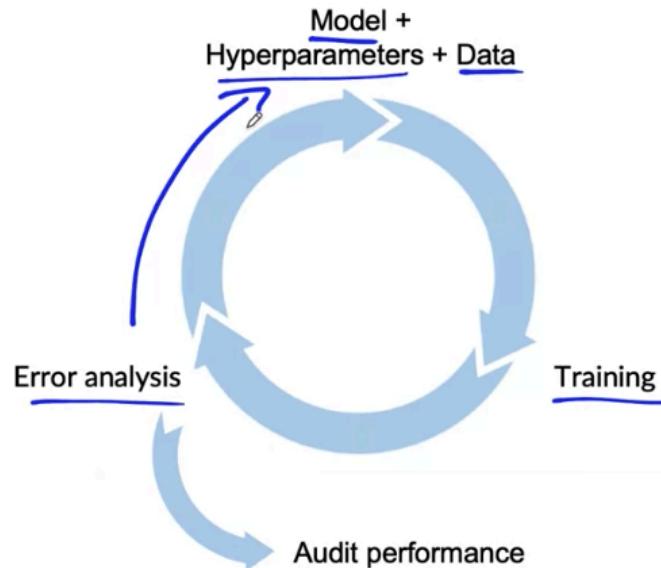
→ 이러한 baseline 설정을 통해 무엇이 가능한지 알 수 있게 되는 것입니다. baseline이 정해진다면 현실적인 프로젝트의 목표치를 정하는 데 많은 도움이 된다고 합니다.

Tips for getting started

- ML is an iterative process
- Getting started on modeling
- Deployment constraints when picking model
- Sanity-check for code and algorithm

모델 선택과 학습을 위한 Tip을 알아보자!

ML is an iterative process



→ 머신러닝은 반복적인 프로세스입니다!
(계속해서 반복되는 부분입니다!)

Tips for getting started

- ML is an iterative process
- Getting started on modeling
- Deployment constraints when picking model
- Sanity-check for code and algorithm

Getting started on modeling

- Literature search to see what's possible (courses, blogs, open-source projects).
- Find open-source implementations if available.
- A reasonable algorithm with good data will often outperform a great algorithm with no so good data.

- 문헌조사를 합니다 (강의, 블로그, open-source project)
- 가능하다면 open-source 구현물을 찾아봅니다
- 완벽한 알고리즘보다는 좋은 데이터가 더 낫습니다

Tips for getting started

- ML is an iterative process
- Getting started on modeling
- Deployment constraints when picking model
- Sanity-check for code and algorithm

Deployment constraints when picking a model

Should you take into account deployment contraints when picking a model?

Yes, if baseline is already established and goal is to build and deploy.

No (or not necessarily), if purpose is to establish a baseline and determine what is possible and might be worth pursuing.

Q. 그렇다면 모델 선택 시 모델을 배포 제약사항은 어떻게 될까요?

"네"와 "아니요" 둘 다 정답이 될 수 있습니다

베이스라인을 설정했다면 모델 배포 제약사항을 고려해야 합니다.

반면 베이스라인을 정하지 않았다면 베이스라인을 정하기 위해 무엇이 가능한지 또는 좋은지 결정하는 단계이므로 배포 제약 사항을 고려할 필요가 없다고 합니다.

Tips for getting started

- ML is an iterative process
- Getting started on modeling
- Deployment constraints when picking model
- Sanity-check for code and algorithm

Sanity-check for code and algorithm

- Try to overfit a small training dataset before training on a large one.
- Example #1: Speech recognition
audio transcript
x → y — — — — —
- Example #2: Image segmentation

- Example #3: Image classification

Sanity-check의 장점은 단 몇 분 또는 심지어 몇 초 만에 하나 또는 소수의 예제에 대한 알고리즘을 훈련 시킬 수 있으며 이를 통해 버그를 훨씬 더 빠르게 찾을 수 있습니다.

What is Sanity-check?

Sanity testing means a basic test to evaluate quickly whether a result or claim of a calculation is true. The main criteria are to rule out the obvious fake ...

2

Error analysis and performance auditing



- Error analysis example
- Prioritizing what to work on
- Skewed datasets
- Performance auditing

Error analysis example

- Iterative process of error analysis
- Useful metrics for each tag

Speech recognition example

각각의 error 상황을 tag 별로 나눠서 특정 tag 별로 개선 여지가 있는지를 판단하는 과정!

Example	Label	Prediction	Car Noise	People Noise
1	"Stir fried lettuce recipe"	"Stir fry lettuce recipe"	✓	
2	"Sweetened coffee"	"Swedish coffee"		✓
3	"Sail away song"	"Sell away song"		✓
4	"Let's catch up"	"Let's ketchup"	✓	✓

Error analysis example

- Iterative process of error analysis
- Useful metrics for each tag

Speech recognition example

Example	Label	Prediction	Car noise	People noise	Low bandwidth
1	"Stir fried lettuce recipe"	"Stir fry lettuce recipe"	1		
2	"Sweetened coffee"	"Swedish coffee"		1	1
3	"Sail away song"	"Sell away some"		1	
4	"Let's catch up"	"Let's ketchup"	1	1	1

성능개선에 도움을 줄 수 있는 Tag들을 Spread Sheet로 정리하면 보다 깔끔하게 지표를 정리 할 수 있습니다
Andrew Ng 교수님께서는 좋은 툴도 있다는 말씀을 해주셨습니다!

Error analysis example

- Iterative process of error analysis
- Useful metrics for each tag

“Training model is iterative process,
Deploying model is iterative process,
error analysis is also Iterative process”

Iterative process of error analysis



← Error Analysis도 결국엔 반복이다!



Visual inspection:

- Specific class labels (scratch, dent, etc.) 
- Image properties (blurry, dark background, light background, reflection, ...)
- Other meta-data: phone model, factory



Product recommendations:

- User demographics
- Product features/category

And the goal of this type of process where you come over tag label.
More data come over tag, is to try to come up with a few categories where
you could productively improve the algorithm such as in our earlier
speech example deciding to work on speech with car noise in the background.

Error analysis example

- Iterative process of error analysis
- Useful metrics for each tag

Useful metrics for each tag

- What fraction of errors has that tag? 12%
- Of all data with that tag, what fraction is misclassified? 18%
- What fraction of all the data has that tag?
- How much room for improvement is there on data with that tag?

Prioritizing what to work on

- Prioritizing what to work on
- Adding/improving data for specific categories

Prioritizing what to work on

Type	Accuracy	Human level performance	Gap to HLP	
Clean Speech	94%	95%	1%	
Car Noise	89%	93%	4%	
People Noise	87%	89%	2%	
Low Bandwidth	70%	70%	0%	

예시를 통해 살펴보도록 하겠습니다

Prioritizing what to work on

- Prioritizing what to work on
- Adding/improving data for specific categories

Prioritizing what to work on

Type	Accuracy	Human level performance	Gap to HLP	% of data
Clean Speech	94%	95%	1%	$60\% \rightarrow 0.6\%$
Car Noise	89%	93%	4%	$4\% \rightarrow 0.16\%$
People Noise	87%	89%	2%	$30\% \rightarrow 0.6\%$
Low Bandwidth	70%	70%	0%	$6\% \rightarrow 0\%$

Q. 전체데이터를 100이라고 가정하자,

우리가 학습시킨 모델과 HLP의 갭을 통해 성능 개선의 여지를 발견한다면 60%, 4%, 30%, 6%의 데이터에 대해서 모델은 얼마만큼의 성능 향상을 할 수 있을까?

A. 각각 0.6%, 0.16%, 0.6%, 0%의 성능 향상을 기대할 수 있습니다!

Skewed datasets

- Examples of skewed datasets
- Confusion Matrix
- Combining precision and recall – F1-score

Examples of skewed datasets



Manufacturing example

99.7% no defect

$y=0$

0.3% defect

$y=1$

`print("0")`
99.7%



Medical Diagnosis example: 99% of patients don't have a disease



Speech Recognition example: In wake word detection, 96.7% of the time
wake word doesn't occur

Skewed datasets

- Examples of skewed datasets
- Confusion Matrix
- Combining precision and recall – F1-score

Confusion matrix: Precision and Recall

		Actual	
		$y=0$	$y=1$
Predicted	$y=0$	905 TN	18 FN
	$y=1$	9 FP	68 TP

$\hookrightarrow 914$ $\hookrightarrow 86$

TN : True Negative

TP : True Positive

FN : False Negative

FP : False Positive

$$\text{Precision} = \frac{TP}{TP+FP} = \frac{68}{68+9} = 88.3\%$$

$$\text{Recall} = \frac{TP}{TP+FN} = \frac{68}{68+18} = 79.1\%$$

Skewed datasets

- Examples of skewed datasets
- Confusion Matrix
- Combining precision and recall – F1-score

Combining precision and recall – F_1 score

	Precision (P)	Recall (R)	F_1
Model 1	88.3	79.1	83.4% ←
Model 2	97.0	<u>7.3</u>	13.6%

$$F_1 = \frac{2}{\frac{1}{P} + \frac{1}{R}}$$

TL;DR

Skewed Dataset과 같이 왜곡도가 심한 데이터로 학습한 모델을 평가하기 위해서는 F1-score를 사용하는 것이 모델 평가에 좋습니다.

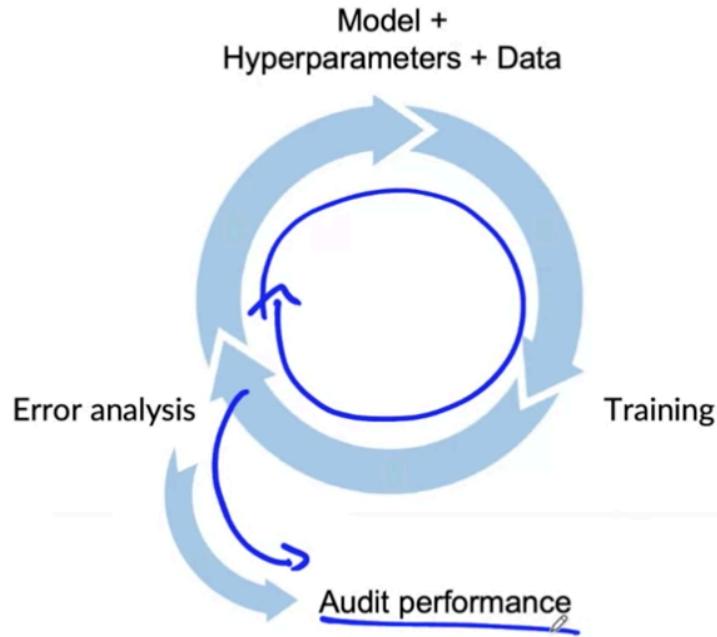
← F1-score는 Precision과 Recall값의 조화 평균으로 구하실 수 있습니다.

Performance auditing

- Performance auditing
- Auditing framework
- e.g Speech recognition example

Q. 모델의 성능은 어떻게 파악할 수 있을까?

Performance auditing



Performance auditing

- Performance auditing
- Auditing framework
- e.g Speech recognition example

“ Accuracy, fairness/biass,
기타 문제점을 체크합니다”

Auditing framework

Check for accuracy, fairness/bias, and other problems.

1. Brainstorm the ways the system might go wrong.

- Performance on subsets of data (e.g., ethnicity, gender).
- How common are certain errors (e.g., FP, FN).
- Performance on rare classes.

← 머리를 맞대고 생각하기



2. Establish metrics to assess performance against these issues on appropriate slices of data.

← Metric 설정하기

3. Get business/product owner buy-in.

← business/product owner

Performance auditing

- Performance auditing
- Auditing framework
- e.g Speech recognition example

Speech recognition example

1. Brainstorm the ways the system might go wrong.

- Accuracy on different genders and ethnicities.
- Accuracy on different devices.
- Prevalence of rude mis-transcriptions.

GAN gun gang

2. Establish metrics to assess performance against these issues on appropriate slices of data.

- Mean accuracy for different genders and major accents.
- Mean accuracy on different devices.
- Check for prevalence of offensive words in the output.

3

Data iteration



- Data-centric AI development
- A useful picture of data augmentation
- Data augmentation
- Can adding data hurt?
- Adding features
- Experiment tracking
- From big data to good data
- Option. Week2. Optional References

Data-centric AI development

Data-centric AI development

Model-centric view

Take the data you have, and develop a model that does as well as possible on it.

Hold the data fixed and iteratively improve the code/model.

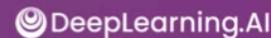
Data-centric view

The quality of the data is paramount. Use tools to improve the data quality; this will allow multiple models to do well.

Hold the code fixed and iteratively improve the data.

→ 데이터의 퀄리티가 중요하다!

→ 모델은 고정한 채로 반복적으로 데이터를 업데이트 시키자!



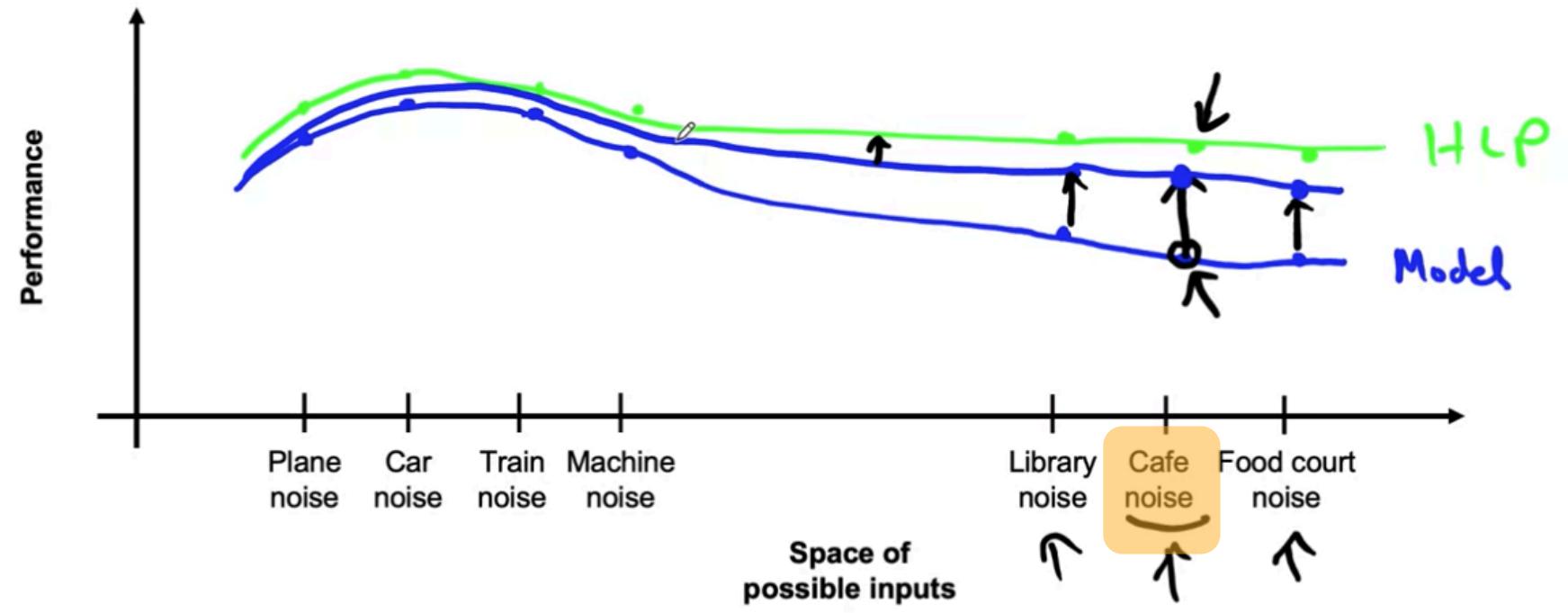
A useful picture of data augmentation

모델의 성능을 파란색, 사람의 성능(HLP)을 초록색으로 표현해보겠습니다

Speech recognition example

Different types of speech input:

- Car noise
- Plane noise
- Train noise
- Machine noise
- Cafe noise
- Library noise
- Food court noise

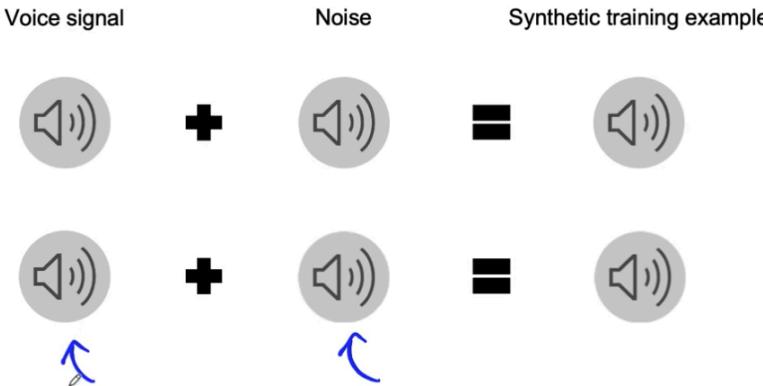


cafe noise에 대한 성능 개선의 여지가 있습니다

→ cafe noise 데이터를 더 수집하여 모델의 성능을 이뤄낼 수 있는 것입니다.

Data augmentation

Data augmentation example



DeepLearning.AI

어떤 종류의 소음(Noise) 일까?
얼마나 시끄러운가?

“Data Augmentation can be a very efficient way to get more data, especially for unstructured data problems such as images, audio, maybe text. But when carrying out data augmentation, there's a lot of choices you have to make. What are the parameters? How do you design the data augmentation setup”

Data augmentation

Data augmentation

Goal:

Create realistic examples that (i) the algorithm does poorly on, but
(ii) humans (or other baseline) do well on



1. 학습 알고리즘이 엄망인 실제 데이터를 만들자
2. 이 때, 인간(HLP)은 데이터 인식을 잘한다

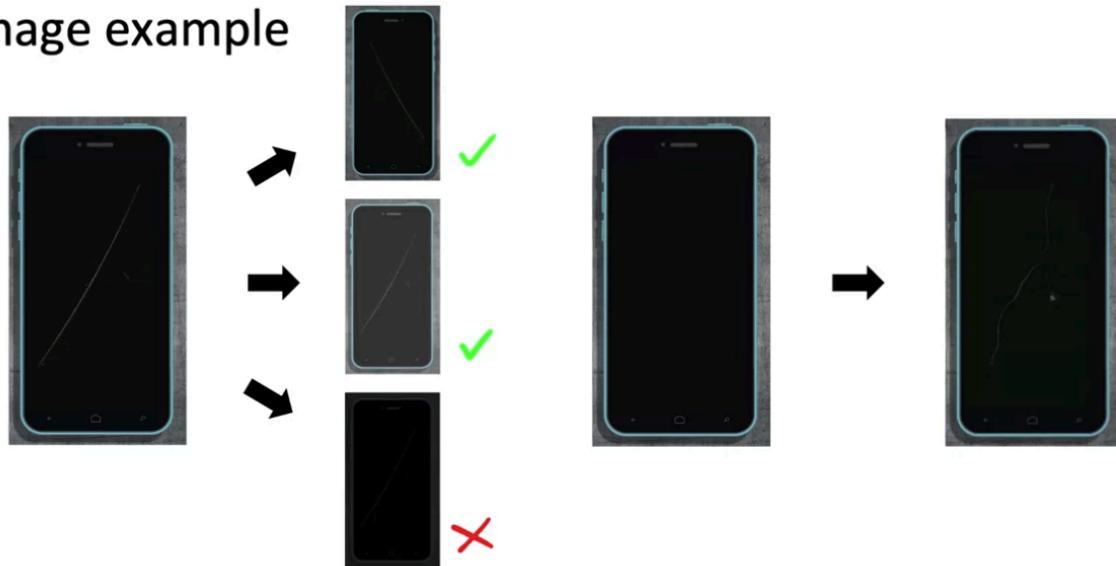
Checklist:

- Does it sound realistic?
- Is the $x \rightarrow y$ mapping clear? (e.g., can humans recognize speech?)
- Is the algorithm currently doing poorly on it?

즉, 인간은 잘 인식하는데 AI System이 잘 인식하지 못하는 데이터를 집중공략해보자!

Data augmentation

Image example



“스크래치가 있는 스마트폰이 있다고 가정해보자!”

사진을 수평선을 기준으로 회전시켜보고,
배경 색을 변경시켜보고...
스마트폰 사진의 스크래치를 포토샵으로 합성해본다

GAN과 같은 복잡한 모델을 사용하여 합성하는 방법도
좋지만 합성이라는 비교적 간단한 방법으로도 Data
Augmentation을 통한 AI System 성능 향상을 경험할 수
있다고 한다!

Can adding data hurt?

Can adding data hurt performance?

For unstructured data problems, if:

- The model is large (low bias).
- The mapping $x \rightarrow y$ is clear (e.g., given only the input x , humans can make accurate predictions).

Then, **adding data rarely hurts accuracy.**

$$P(x) \xrightarrow{\text{20% noise}} \text{cate}$$

"if you're using data augmentation, you're adding to specific parts of the training set such as adding lots of data with cafe noise... is this going to hurt your learning album's performance?"

Q. Training Data 의 Specific part에 대해서 더이터 증분을 진행할 경우, 모델의 성능에 어떤 영향을 미칠까?

Can adding data hurt?

Can adding data hurt performance?

For unstructured data problems, if:

- The model is large (low bias).
- The mapping $x \rightarrow y$ is clear (e.g., given only the input x , humans can make accurate predictions).

Then, **adding data rarely hurts accuracy.**

$$\begin{array}{c} P(x) \\ \xrightarrow{\quad\quad\quad} \\ 20\% \text{ Cafe} \\ \downarrow \\ 50\% \end{array}$$

비정형 데이터에 대해...

- 모델이 복잡하고(bias가 낮고)
- 레이블링이 잘 되어 있으면 ($x \rightarrow y$)
- Data를 더해주는 것이 정확도를 낮추지 않는다

"if you're using data augmentation, you're adding to specific parts of the training set such as adding lots of data with cafe noise... is this going to hurt your learning album's performance?"

because adding data through data augmentation or collecting more of one type of data, **can really change your input dated distribution to probability of x.**

Let's say at the start of your problem, 20% of your data had cafe noise. But using augmentation, you added a lot of cafe noise. So now this is 50 of your data is data of cafe noise in the background.

this could hurt the performance on non cafe noise data.
But if your model is large enough, then this isn't really an issue.

이러한 점은 주의하자

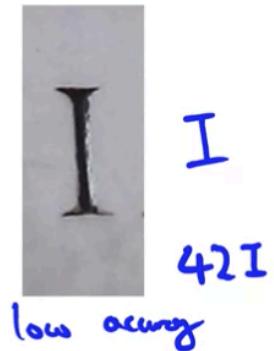
- 데이터가 작은 경우**
- Rare class의 데이터를 더할 경우 데이터셋 분포에 영향을 줄 수 있다
- 이 경우는 문제가 될 수 있다**

Can adding data hurt?

Photo OCR counterexample



1



I I

42I
low accuracy

1? I?

Adding a lot of new "I"s may skew the dataset and hurt performance → 이 경우는 문제가 될 수 있다

DeepLearning.AI

'I'를 너무 많이 더해버렸어... 데이터 분포가 바뀌었잖아!

Adding features

Structured data



Restaurant recommendation example

Vegetarians are frequently recommended restaurants with only meat options.

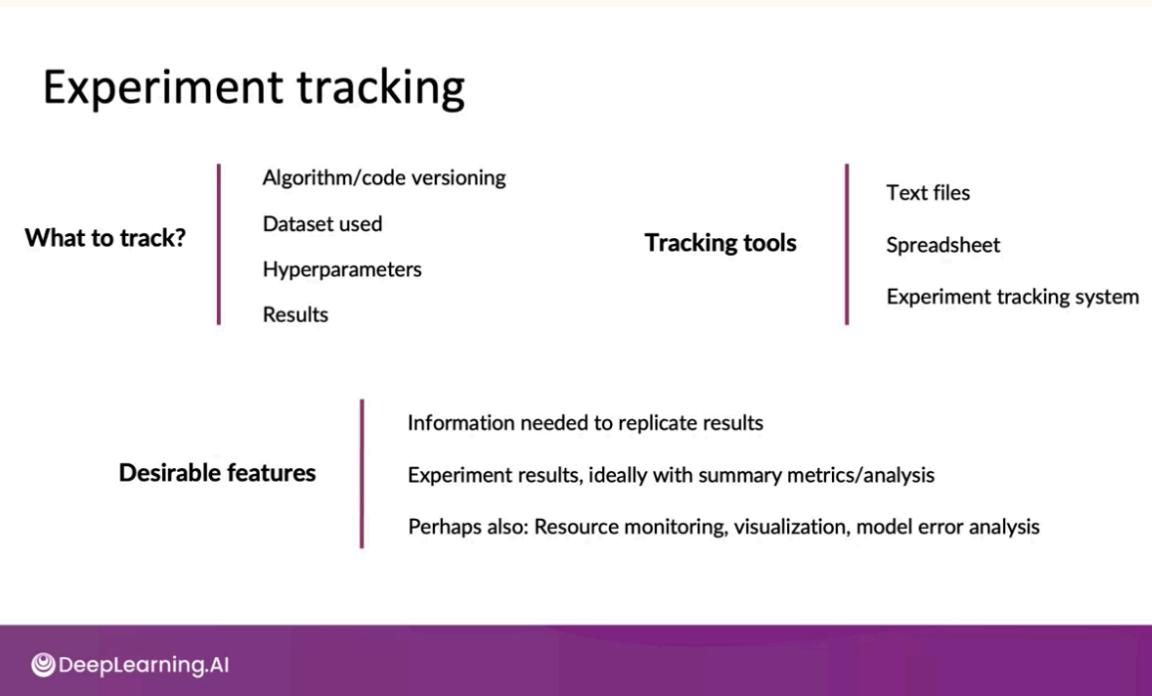


Possible features to add?

- Is person vegetarian (based on past orders)?
- Does restaurant have vegetarian options (based on menu)?

Q. 어떤 feature 값을 더해주면 모델이 조금 더 나은 판단을 할 수 있을까?

Experiment tracking

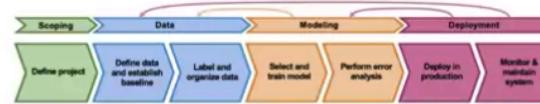


- **What to track?: 어떤 정보를 추적할 것인가?**
 - Algorithm/code versioning
 - Dataset used
 - Hyper parameters
 - Results
- **Tracking tools: 어떤 도구를 사용할 것인가?**
 - Text files
 - Spreadsheet
 - Exp tracking system
- **Desirable features: 적합한 특징값은 무엇일까?**
 - Information need to replicate results
 - Experiment results
 - Resource monitoring, visualization, model error analysis

From big data to good data

From Big Data to Good Data

Try to ensure consistently high-quality data in all phases of the ML project lifecycle.



◀ 머신러닝 프로젝트 개발과정에서
양질의 데이터를 보장해주세요!

Good data:

- Covers important cases (good coverage of inputs x)
- Is defined consistently (definition of labels y is unambiguous)
- Has timely feedback from production data (distribution covers data drift and concept drift)
- Is sized appropriately

Q. 좋은 데이터란?

- 중요한 모든 경우를 커버함
- 명확한 레이블링(y)
- 주기적으로 Data Drift, Concept Drift에 대한 피드백
- 적절한 사이즈!

Review

- 모델 개발의 주요 과제를 학습했습니다.
- 불균형적으로 중요한 소수의 예에 대한 성능이 대다수의 예에 대한 성능보다 더 중요할 수 있는 방법을 설명합니다.
- 훈련 데이터의 드문 클래스가 성능에 어떤 영향을 미칠 수 있는지 알아보았습니다.
- 모델 퍼포먼스에 대한 기준(baseline for your performance)을 설정하는 세 가지 방법을 정의했습니다.
- 정형 데이터와 비정형 데이터를 정의했습니다.
- 모델을 선택할 때 배포 제약 조건을 고려해야 하는 경우를 확인하였습니다.



Review

- ML 모델링 시작과 관련된 단계를 알아보았습니다.
- 오류 분석을 위한 반복 프로세스를 설명했습니다.
- 모델 정확도를 개선하기 위해 작업할 때 우선 순위를 정할 항목을 결정할 때 핵심 요소를 식별하였습니다.
- 오디오 데이터와 이미지 데이터를 비교하여 데이터 증대에 사용할 수 있는 방법을 설명하였습니다.
- 심하게 치우친 데이터 세트에 대해 학습할 수 있는 문제를 설명하였습니다.
- 훈련 데이터 세트에 더 많은 데이터를 추가하면 실제로 성능이 저하될 수 있는 사용 사례를 식별했습니다.
- 실험 추적의 주요 구성 요소를 설명하였습니다.





끝



Q.

질문 있으신가요?



함께 묻고 답하는 MLOps 스터디가 되었으면 하는 바램입니다

- Error Analysis를 위한 Tagging이란 무엇인가요? - ppt23
- 실제 현업에서 데이터 관리는 어떻게 하고 있는가?
- 데이터와 모델 관리는 어떻게 하고 있나요?
 - 데이터의 품질을 검증하는 방법은 어떤 방법이 있을까요?
- 수집한 데이터에 대한 레이블, 태깅은 어떻게 하고 있나요? 어떻게 하면 좋을까요?
- Computer Vision 문제에서 데이터의 분포를 파악하는 것이 데이터를 파악하는 방법일까요?
다른 방법은 없을까요?

감사합니다!