# Course Overview

# INDEX

1. Syllabus

2. Syllabus - Project

3. Data Scientist

4. AI

# Syllabus

**정형**

- 데이터 분석 환경 구축
- 파이썬 기초(Numpy, Pandas)
- 시각화 라이브러리
- 데이터의 결측
- 데이터의 이상치

- 데이터 특성 및 이상치 해결
- 탐색적 데이터 분석
- 데이터 시각화
- 상관관계
- 스케일링

- Random Forest
- LighgtGBM
- 앙상블

- Confusion Matrix
- Classification
- Sampling의 종류와 기술
- AUC-ROC curve
- 전처리 기초(결측치, 중복처리)
- 이상치 탐지와 처리

- 범주형 처리
- 스케일링과 정규화
- 특성(피처) 엔지니어링

# Syllabus

## 시계열

- 웹 크롤링
- 데이터 전처리
- 데이터 시각화

- 단층, 다층 신경망 적용
- 단층, 다층 신경망의 이해
- CNN
- RNN
- LSTM

- 시계열 데이터 수집
- 데이터 전처리

- DL 모델 적용
- FNN
- CNN, RNN, AE
- 평가지표
- 결과 시각화
- 모델 해석 도구, 적용

# Syllabus

## 비정형

- 이미지 데이터 수집
- 이미지 데이터 라벨링

- AlexNET
- VGG
- GoogLe Net
- RCNN
- mAP

- OD – SSD, YOLO, R-CNN

- LM

- ChatGPT, Open API
- Llama
- Prompt Engineering
- LangChain
- privateGPT
- LLM Service (API)

# Syllabus

✓ **Skill Set**

**Python**

Python 기초 프로그래밍, numpy, pandas, openCV, matplotlib

**Machine Learning Model**

Linear Regression, Logistic Regression, Decision Tree, Random Forest, PCA, Support Vector Machine, Xgboost, Catboost

**Deep Learning Model**

FNN, RNN, LSTM, CNN, VGG, Resnet, Seq2Seq, Transformer

**Data skill**

정형 데이터, 시계열 데이터(시간 의존성, 주기성, 변동성), 비정형 데이터

**LLM(거대언어모델)**

Natural Language Processing(NLP) 생성모델 이해, LLM 이해, Gemini API, Prompt Engineering, PrivateGPT, LLM을 활용한 서비스 예제

# Syllabus

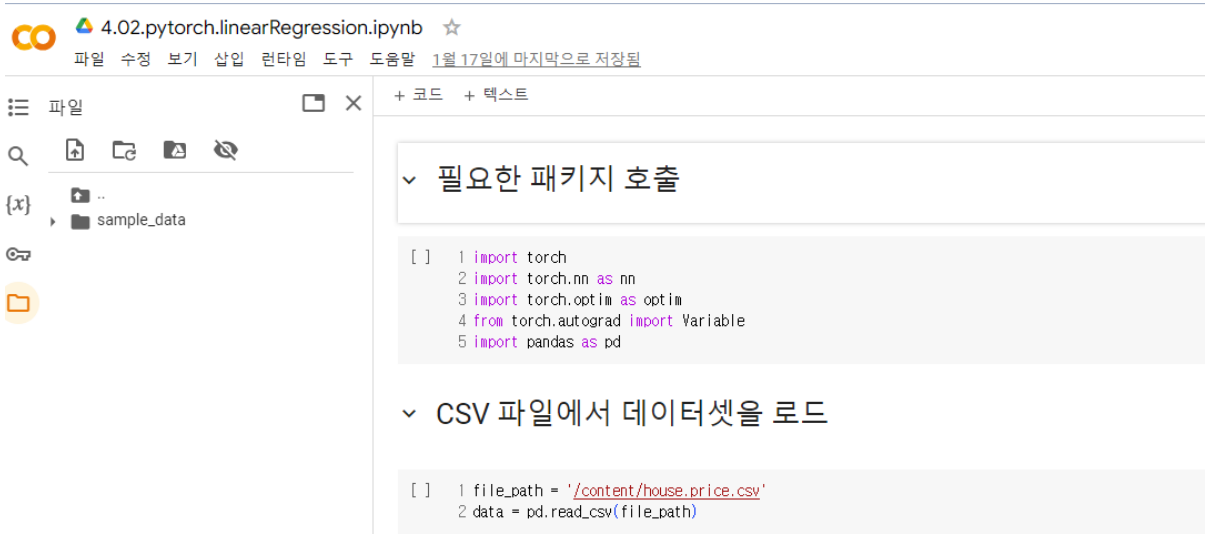
## ✓ Daily TimeTable

| 구분 | 수업시간 | 수업내용 |
|---|---|---|
| 1교시 | 09:00 ~ 09:50 | 강사 수업 |
| 2교시 | 10:00 ~ 10:50 | |
| 3교시 | 11:00 ~ 11:50 | |
| 점심시간 | 12:00 ~ 12:50 | |
| 4교시 | 13:00 ~ 13:50 | |
| 5교시 | 14:00 ~ 14:50 | |
| 6교시 | 15:00 ~ 15:50 | |
| 7교시 | 16:00 ~ 16:50 | 멘토님들 회고 시간<br>+Daily / Weekly Quiz |
| 8교시 | 17:00 ~ 17:50 | |
| 퇴실 | 17:50 ~ 18:00 | 강의 종료 |

# Syllabus

## ✓ 구글 코랩, 주피터 노트북

| 특징 | 구글 코랩 | 주피터 노트북 |
|---|---|---|
| **설치** | 클라우드 기반이므로 설치 불필요 | 내 컴퓨터에 설치 |
| **실행 환경** | 인터넷이 연결된 웹 브라우저 | 오프라인 상태의 웹 브라우저 |
| **파일 저장** | 구글 드라이브에 자동 저장 | 내 컴퓨터 하드 디스크에 저장 |
| **패키지** | 사전 설치되어 제공 | 패키지 관리자로 직접 설치 |



# Syllabus – Project

# 아마존 리뷰 요약

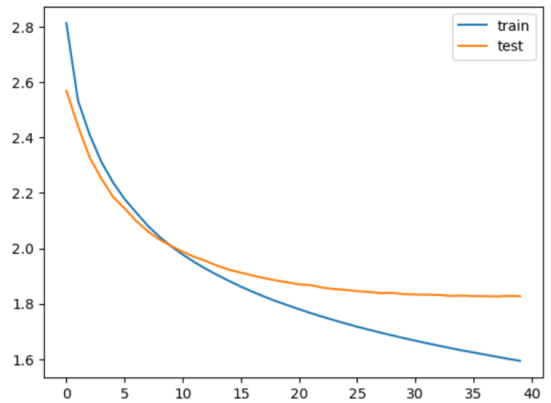
## 모델 학습

### Sparse categorical crossentropy

- 주로 다중 클래스 분류 문제에서 사용
- 실제 레이블이 정수 형태로 주어짐
- 각 샘플의 예측 확률 분포와 실제 레이블 간의 차이를 측정

```
Epoch 37/50
914/914 [==============================] - 344s 377ms/step - loss: 1.6161 - val_loss: 1.8271
Epoch 38/50
914/914 [==============================] - 349s 382ms/step - loss: 1.6084 - val_loss: 1.8261
Epoch 39/50
914/914 [==============================] - 347s 380ms/step - loss: 1.6008 - val_loss: 1.8280
Epoch 40/50
914/914 [==============================] - 353s 386ms/step - loss: 1.5937 - val_loss: 1.8269
Epoch 40: early stopping
```

6시간 소요




## 테스트 결과

### Input review

I am really happy with this oatmeal, tastes the same as the regular stuff, without the extra sugar. Also, the 'lower sugar' is hard to find at my local grocer, and when they do have it, the price is almost twice what Amazon is charging. Granted, I have to buy 4 boxes at a time, compared to one, but that's what my pantry is for.

이 오트밀로 정말 만족해. 일반적인 것과 똑같이 맛도 나고, 추가 설탕도 없어. 게다가, '낮은 설탕 함량' 제품을 내 현지 식료품점에서 찾기가 힘들어. 그리고 그들이 있을 때는 아마존이 청구하는 가격의 거의 두 배나 돼. 물론, 한 번에 4상자씩 사야 돼지만, 그게 내 식료품 보관함이 있는 곳이잖아.

### LSTM Decoder

great product but overpriced

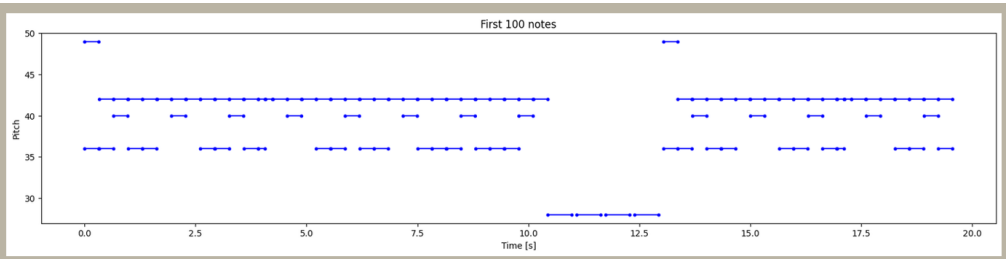
### Attention Decoder

great oatmeal cheaper than grocery store

# Syllabus – Project

# A IYRICS

## AI를 활용한 작사 작곡 모델




악기가 끊임없이 연주되는 것들을 멜로디라고 가정 후, 멜로디 따로 추출

멜로디 아닌 다른 베이스 악기들도 따로 저장

### 원본 가사

**Last Chance - 소수빈**

I knew that

넌 매순간, 영화처럼 곁에

기적처럼 앞에

그렇게 넌 내게 와

Every time and everywhere

그댈보며 난 웃고

널 그리며, 떠올려

너와 같은 온도로, 떨림으로

난 더 가까이

발을 옮겨, 다시

### LSTM_bigram

i m still i m here

넌 나의 <unk>

기적처럼

그렇게 <unk>

every day i m so lucky 숨겨왔던

<unk> <unk>

너와 함께 했던 그 말

난 <unk>

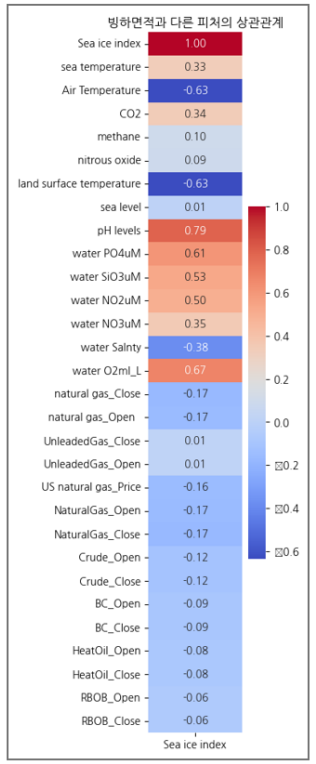
발을 <unk>

# Syllabus – Project

## 해빙 예측 모델

### -줄어드는 빙하

| 모델 | R^2 | Rank | 모델 | MSE | Rank | 모델 | MAE | Rank | 모델 | RMSE | Rank |
|---|---|---|---|---|---|---|---|---|---|---|---|
| new_lstm | 0.9965 | 1 | new_lstm | 0.0423 | 1 | new_lstm | 0.1568 | 1 | new_lstm | 0.2058 | 1 |
| SVM | 0.9947 | 2 | SVM | 0.0661 | 2 | KNN | 0.1706 | 2 | SVM | 0.2571 | 2 |
| KNN | 0.9941 | 3 | KNN | 0.0737 | 3 | SVM | 0.1816 | 3 | KNN | 0.2715 | 3 |
| Random Forest | 0.9936 | 4 | Random Forest | 0.0789 | 4 | Random Forest | 0.1924 | 4 | Random Forest | 0.2809 | 4 |
| XGBoost | 0.9928 | 5 | XGBoost | 0.0893 | 5 | XGBoost | 0.208 | 5 | XGBoost | 0.2988 | 5 |
| LSTM | 0.9913 | 6 | LSTM | 0.1075 | 6 | Decision Tree | 0.2364 | 6 | LSTM | 0.3279 | 6 |



| | | | | | |
|---|---|---|---|---|---|
| Air Temperature | 0.225488 | water NO3uM | 0.120678 | methane | 0.109495 |
| sea temperature | 0.211735 | water Salnty | 0.119471 | Crude_Close | 0.109238 |
| water PO4uM | 0.151907 | NaturalGas_Open | 0.117754 | RBOB_Close | 0.108882 |
| water O2ml_L | 0.145461 | natural gas_Open | 0.117462 | HeatOil_Open | 0.108648 |
| pH levels | 0.137823 | HeatOil_Close | 0.115482 | BC_Close | 0.108358 |
| water SiO3uM | 0.137506 | natural gas_Close | 0.11366 | sea level | 0.108048 |
| land surface temperature | 0.130546 | RBOB_Open | 0.11284 | water NO2uM | 0.107256 |
| UnleadedGas_Open | 0.127031 | NaturalGas_Close | 0.111928 | BC_Open | 0.107017 |
| UnleadedGas_Close | 0.126732 | CO2 | 0.111183 | nitrous oxide | 0.104789 |
| US natural gas_Price | 0.124997 | Crude_Open | 0.110322 | | |

# Syllabus – Project

웹 크롤링 리뷰 데이터를
활용한 약 이름 예측

01

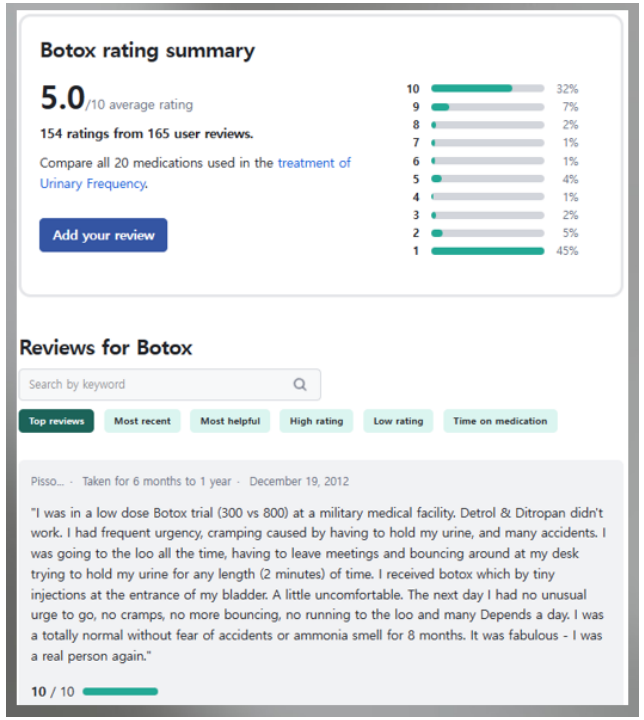
리뷰와 증상을 활용한
약의 이름과 카테고리 예측

웹 크롤링 리뷰 데이터(drugs.com)를 활용하여
리뷰, 증상 컬럼을 주된 입력으로 두고,
약의 이름과 카테고리를 예측

02

리뷰 키워드를 활용하여 증상
을 예측하고 알맞은 약의 이름
출력

리뷰의 키워드를 통해 증상을 예측하고,
증상에 맞는 약의 이름을 출력




방법론 및 모델링 01
모델링 성능

/27

전체 Drug Name
( 159958 rows )
( Target : 3429 )
Accuracy : 0.2863
Precision : 0.2302
Recall : 0.2863
F1 Score : 0.2198

| 75342 rows | 52508 rows | 39482 rows | 21049 rows | 9671 rows |
|---|---|---|---|---|
| TOP 100 Drug Name ( Target : 100 ) | TOP 50 Drug Name ( Target : 50 ) | TOP 30 Drug Name ( Target : 30 ) | TOP 10 Drug Name ( Target : 10 ) | TOP 3 Drug Name ( Target : 3 ) |
| Accuracy : 0.4089 | Accuracy : 0.4931 | Accuracy : 0.5246 | Accuracy : 0.6803 | Accuracy : 0.9044 |
| Precision : 0.3969 | Precision : 0.4903 | Precision : 0.5033 | Precision : 0.6573 | Precision : 0.9087 |
| Recall : 0.4089 | Recall : 0.4931 | Recall : 0.5245 | Recall : 0.6803 | Recall : 0.9044 |
| F1 Score : 0.3755 | F1 Score : 0.4656 | F1 Score : 0.4880 | F1 Score : 0.6400 | F1 Score : 0.9049 |
| Category of TOP 100 DrugName ( Target : 15 ) | Category of TOP 50 DrugName ( Target : 12 ) | Category of TOP 30 DrugName ( Target : 9 ) | Category of TOP 10 DrugName ( Target : 3 ) | Category of TOP 3 DrugName ( Target : 1 ) |
| Accuracy : 0.7904 | Accuracy : 0.8476 | Accuracy : 0.9119 | Accuracy : 0.9914 | Accuracy : 1 |
| Precision : 0.7941 | Precision : 0.8492 | Precision : 0.9149 | Precision : 0.9915 | Precision : 1 |
| Recall : 0.7904 | Recall : 0.8476 | Recall : 0.9119 | Recall : 0.9914 | Recall : 1 |
| F1 Score : 0.7789 | F1 Score : 0.8341 | F1 Score : 0.9035 | F1 Score : 0.9914 | F1 Score : 1 |

# Syllabus – Project

## 모델기반 질병 예측 및 약물 추천 테스트 결과

Predicted Condtion: 리뷰를 통해 예측된 질병

Answer Condition: 리뷰에 해당하는 정답 질병

Recommeded Drugs based on ratings: 예측된 질병의 평점 상위 5개 약물 추천 결과

```
1/1 ━━━━━━━━━━━━━━━━━━━━ 0s 25ms/step
Predicted Condition: Birth Control
Answer Condition: Birth Control
Recommended Drugs based on ratings only:
1 Plan B
2 Heather
3 Elinest
4 Provera
5 Ortho-Novum 7 / 7 / 7
1/1 ━━━━━━━━━━━━━━━━━━━━ 0s 26ms/step
Predicted Condition: Bipolar Disorde
Answer Condition: Bipolar Disorde
Recommended Drugs based on ratings only:
1 Eskalith-CR
2 Nuvigil
3 Armodafinil
4 Klonopin Wafer
5 Tiagabine
```

# Syllabus – Project

시퀀스데이터 분석 프로젝트

## 이미지를 통한 알약 분석

### COLUMN INFORMATION

(Row: 25,125 / Column: 30)

| # | Column | Dtype |
|---|---|---|
| 1 | 품목일련번호 | int64 |
| 2 | 품목명 | object |
| 3 | 업소일련번호 | int64 |
| 4 | 업소명 | object |
| 5 | 성상 | object |
| 6 | 큰제품이미지 | object |
| 7 | 표시앞 | object |
| 8 | 표시뒤 | object |
| 9 | 의약품제형 | object |
| 10 | 색상앞 | object |

| # | Column | Dtype |
|---|---|---|
| 11 | 색상뒤 | object |
| 12 | 분할선앞 | obect |
| 13 | 분할선뒤 | int64 |
| 14 | 크기장축 | object |
| 15 | 크기단축 | object |
| 16 | 크기두께 | object |
| 17 | 이미지생성일자 | int64 |
| 18 | 분류번호 | object |
| 19 | 분류명 | object |
| 20 | 전문일반구분 | object |

| # | Column | Dtype |
|---|---|---|
| 21 | 품목허가일자 | int64 |
| 22 | 제형코드명 | obect |
| 23 | 표기내용앞 | int64 |
| 24 | 표기내용뒤 | object |
| 25 | 표기이미지앞 | object |
| 26 | 표기이미지뒤 | object |
| 27 | 표기코드앞 | object |
| 28 | 표기코드뒤 | object |
| 29 | 변경일자 | object |
| 30 | 사업자번호 | int64 |

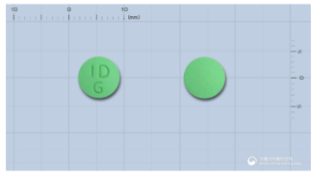
### COLOR(BACK)

| 색상앞 | 개수 |
|---|---|
| 하양\|투명 | 24 |
| 분홍\|투명 | 17 |
| 노랑\|옅은 | 17 |
| 주황\|투명 | 14 |
| 파랑\|옅은 | 14 |
| 하양\|파랑 | 12 |
| 하양\|노랑 | 8 |
| 갈색\|옅은 | 5 |
| 갈색\|진한 | 4 |
| 노랑\|투명 | 4 |
| 회색\|진한 | 4 |
| 초록\|투명 | 2 |
| 청록\|투명 | 2 |
| 분홍\|옅은 | 1 |

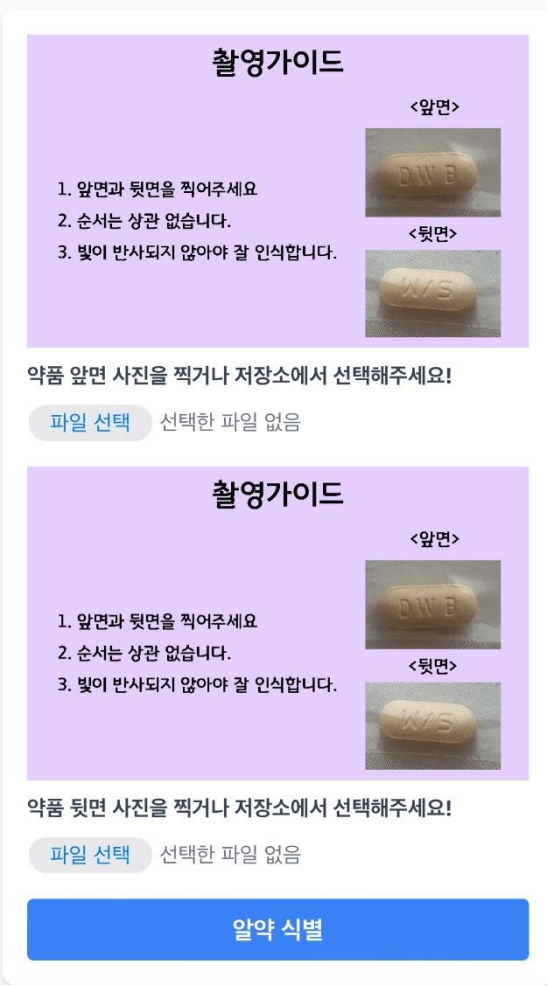
| 색상앞 | 개수 |
|---|---|
| 노랑\|진한 | 1 |
| 초록\|진한 | 1 |
| 초록\|옅은 | 1 |
| 검정\|투명 | 1 |
| 청록\|옅은 | 1 |
| 파랑\|진한 | 1 |
| 주황\|옅은 | 1 |
| 자주\|진한 | 1 |
| 하양\|청록 | 1 |
| 하양\|초록\|투명 | 1 |
| 자주\|옅은 | 1 |
| 연두\|진한 | 1 |
| - | 21168 |

두 개 이상의 색상 값을 갖는 행 140개
→ 가장 앞의 1개 색상 값만 사용

'-' 값을 갖는 행 1개
→ 앞면 색상으로 각각 대체



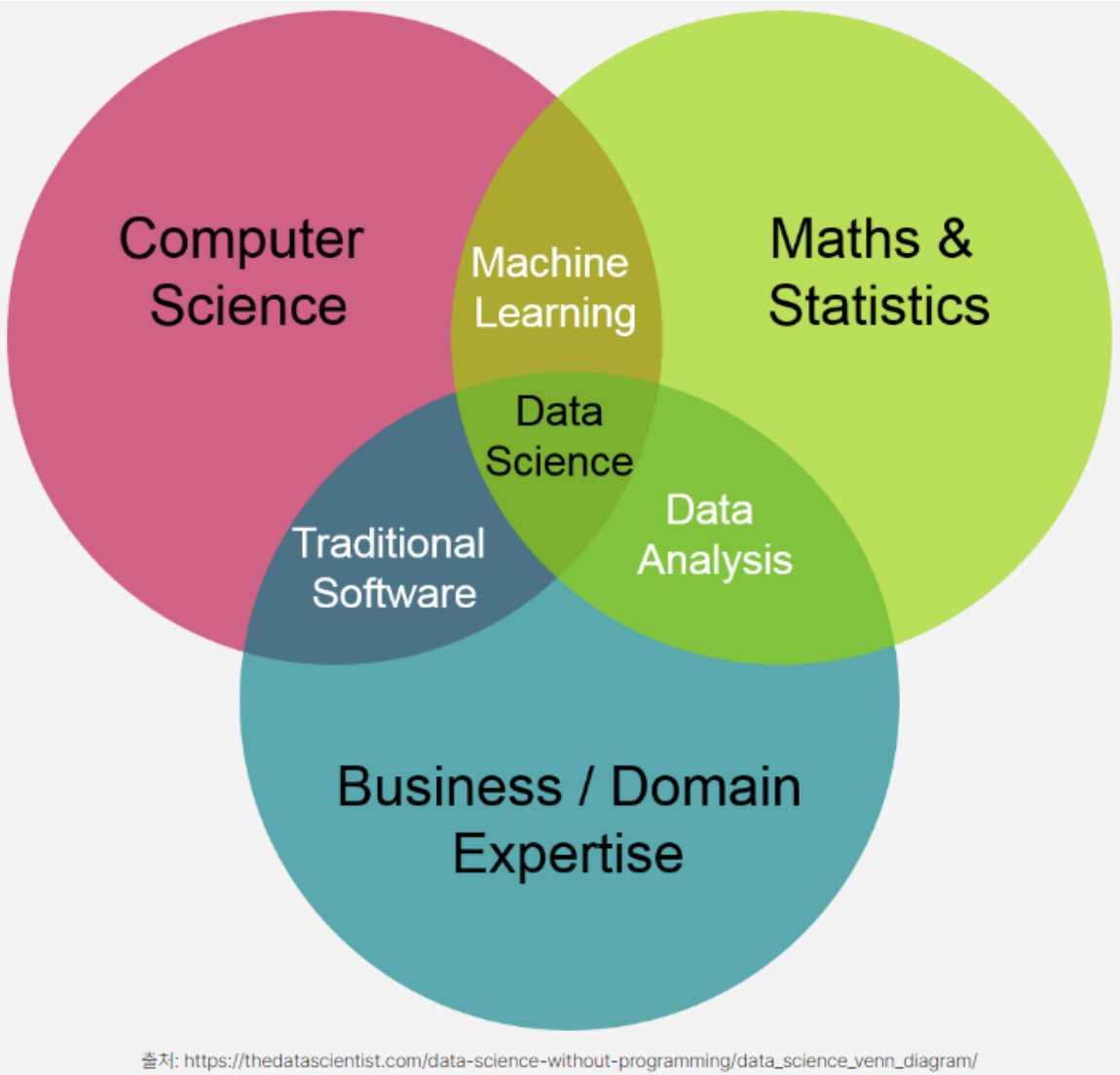
| | 1번 모델 | 2번 모델 | 3번 모델 |
|---|---|---|---|
| 구조 | Conv2D(32) → MaxPooling2D → Conv2D(64) → MaxPooling2D → Flatten → Dense(512) → Dense(256) → Dense(classes) | Conv2D(32) → MaxPooling2D → Conv2D(128) → MaxPooling2D → Flatten → Dense(512) → Dense(256) → Dense(classes) | Conv2D(32) → MaxPooling2D → Conv2D(64) → Dropout → MaxPooling2D → Flatten → Dense(512) → Dense(256) → Dense(classes) |
| 특징 | 두 개의 Conv2D 층과 두 개의 완전 연결된 층 | 1번 모델보다 두번째 Conv2D층의 필터수가 많음 | 첫 번째 Conv2D층과 두 번째 Conv2D층 사이에 Dropout 추가 → 과적합 방지 |
| 성능 | Test loss : 0.348390<br>Test accuracy : 0.894513 | Test loss : 0.775313<br>Test accuracy : 0.810004 | Test loss : 0.451353<br>Test accuracy : 0.852964 |



# Data Scientist

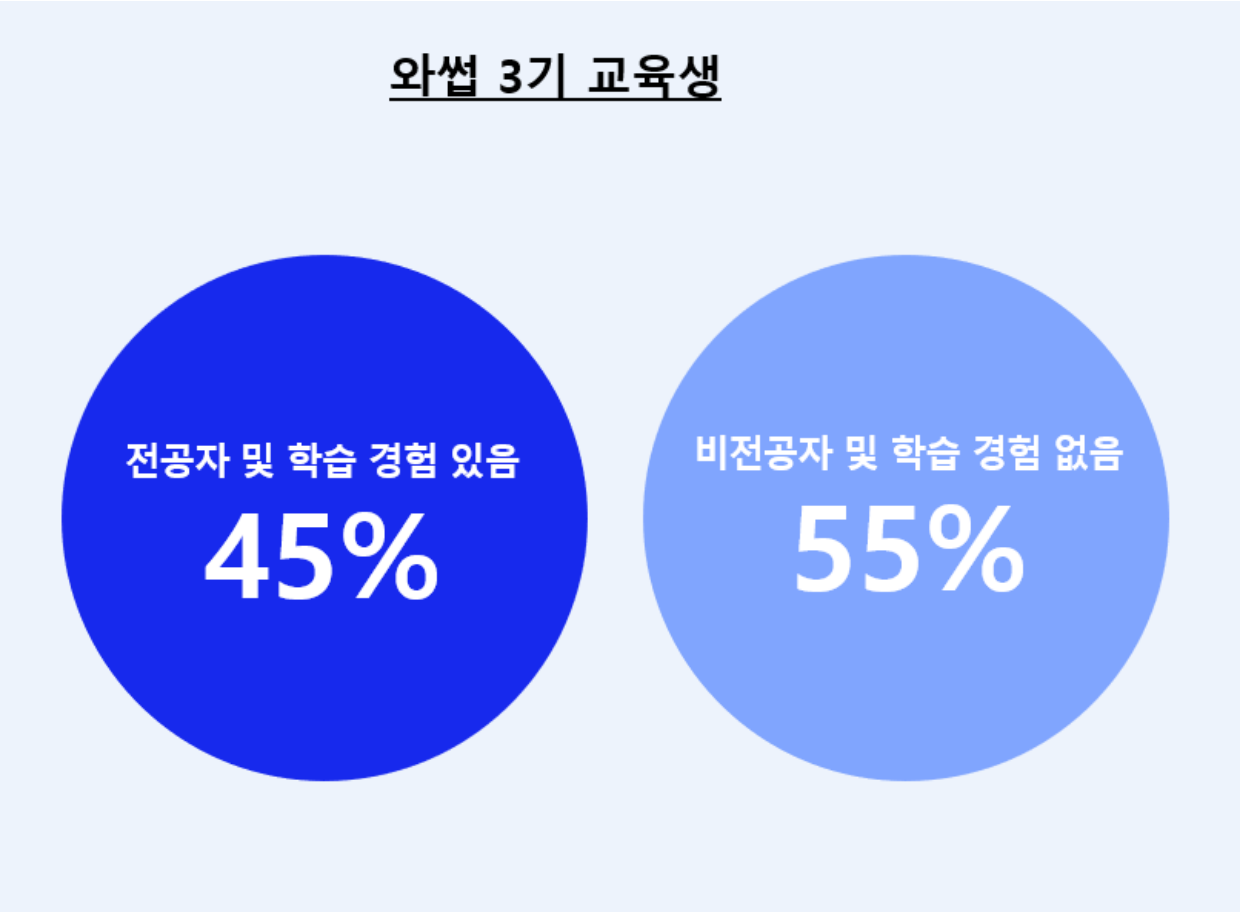
## Data Scientist

Data scientists are analytical experts who extract meaning from and interpret data to solve complex problems. They use industry knowledge, contextual understanding and skepticism of existing assumptions to uncover solutions to business challenges.

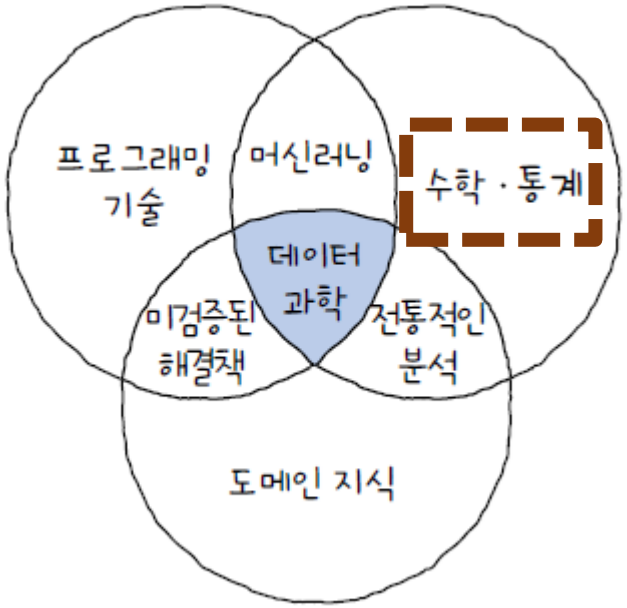



출처: https://thedatascientist.com/data-science-without-programming/data_science_venn_diagram/

# Data Scientist



# Data Scientist




## 어느 정도까지 이해 ???

- 가중치 업데이트 (cont'd)
  - ② $\frac{\partial \hat{y}_i}{\partial Z_{out}}$, $\hat{y} = \sigma(Z_{out}) = \frac{1}{1+e^{-Z_{out}}}$
    - How to differentiate?
    - Then, we get
      - $\frac{\partial \hat{y}_i}{\partial Z_{out}} = \frac{e^{-Z_{out}}}{(1+e^{-Z_{out}})^2} = \frac{1}{1+e^{-Z_{out}}}\left(1 - \frac{1}{1+e^{-Z_{out}}}\right) = \hat{y}_i(1 - \hat{y}_i$

# Data Scientist

✓ Data

- 인공지능의 연료
- <u>센싱기술의 발달</u>
- 빅데이터

# Data Scientist - Role

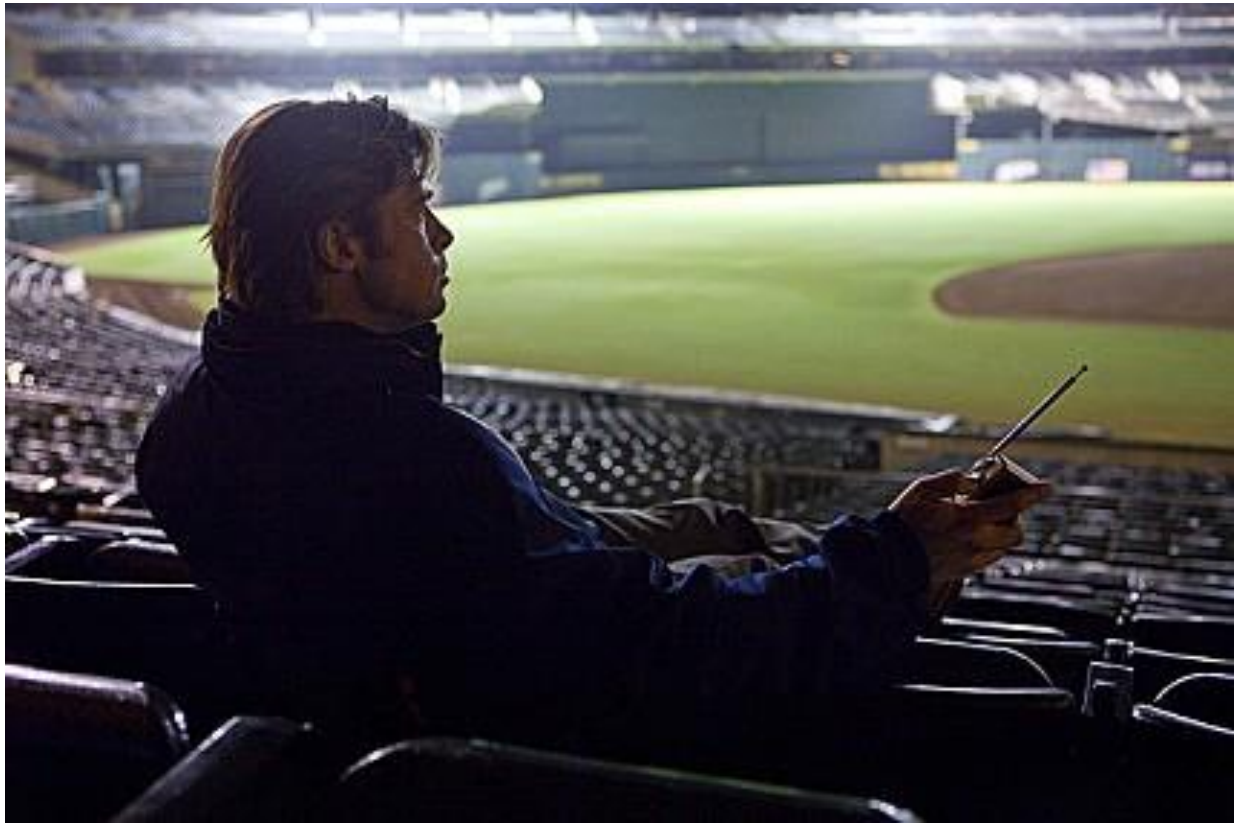
- 데이터 분석가 (Data Analyst)
  데이터를 분석, 비즈니스 결정에 필요한 인사이트 도출.
- 머신러닝 엔지니어 (Machine Learning Engineer)
  머신러닝 모델을 설계, 개발, 배포하고 유지보수, 시스템에 모델 통합
- 데이터 사이언티스트 (Data Scientist)
  데이터 분석과 머신러닝 기술을 활용, 문제 해결.  예측 모델링과 통계 분석 수행
- AI 연구원 (AI Researcher)
  새로운 알고리즘, 모델을 연구/개발
- 딥러닝 엔지니어 (Deep Learning Engineer)
  딥러닝 모델을 전문적으로 다루며, 특히 이미지 인식, 자연어 처리 등 수행

# Data Scientist - Role

## 2024 노벨화학상 '단백질 분야' 공헌 생화학자·AI 전문가 3명 수상

# Data Scientist - Role

## 영화 – 머니볼 (2011)



# Data Scientist - Role

- **'프롬프트 엔지니어'**
  - ✓ AI 스타트업 안트로픽(Anthropic). 프롬프트 엔지니어 채용 공고 '28~37만 달러의 높은 연봉'
  - ✓ 온라인 뉴스 미디어 보딩에어리어(BoardingArea). 시간제 챗GPT 전문가 채용
  - ✓ 프리랜서 플랫폼 업워크(Upwork). 시간당 최대 40달러에 프롬프트 엔지니어를 고용 (웹사이트 콘텐츠 제작)
  - ✓ AI 기반 계약 검토 업체 클라리티(Klarity). 프롬프트 엔지니어 구인 '23만 달러 연봉'.
  - ✓ AI 스타트업 뤼튼테크놀로지스. 프롬프트 엔지니어를 공개 채용 '최대 1억 원의 연봉'
    - 뤼튼 서비스에 사용되는 다양한 목적의 프롬프트 제공, 테스트, 문서화
    - 다양한 생성형 AI 모델을 프롬프트 관점에서 분석 및 정리, 프롬프트 모범 사례 구축 등 업무 수행

- **마켓플레이스** (자신만의 프롬프트를 무료로 공유하거나 사고파는 플랫폼)
  - ✓ 프롬프트베이스(PromptBase)다.
    - 챗GPT(ChatGPT), 스테이블 디퓨전(Stable Diffustion), 미드저니(Midjourney) 등의 프롬프트 거래
    - 1.99달러에서 9.99달러 사이의 가격 (판매 수익의 20% 수수료 부과).

# Data Scientist - Role

https://v.daum.net/v/20250220113710958

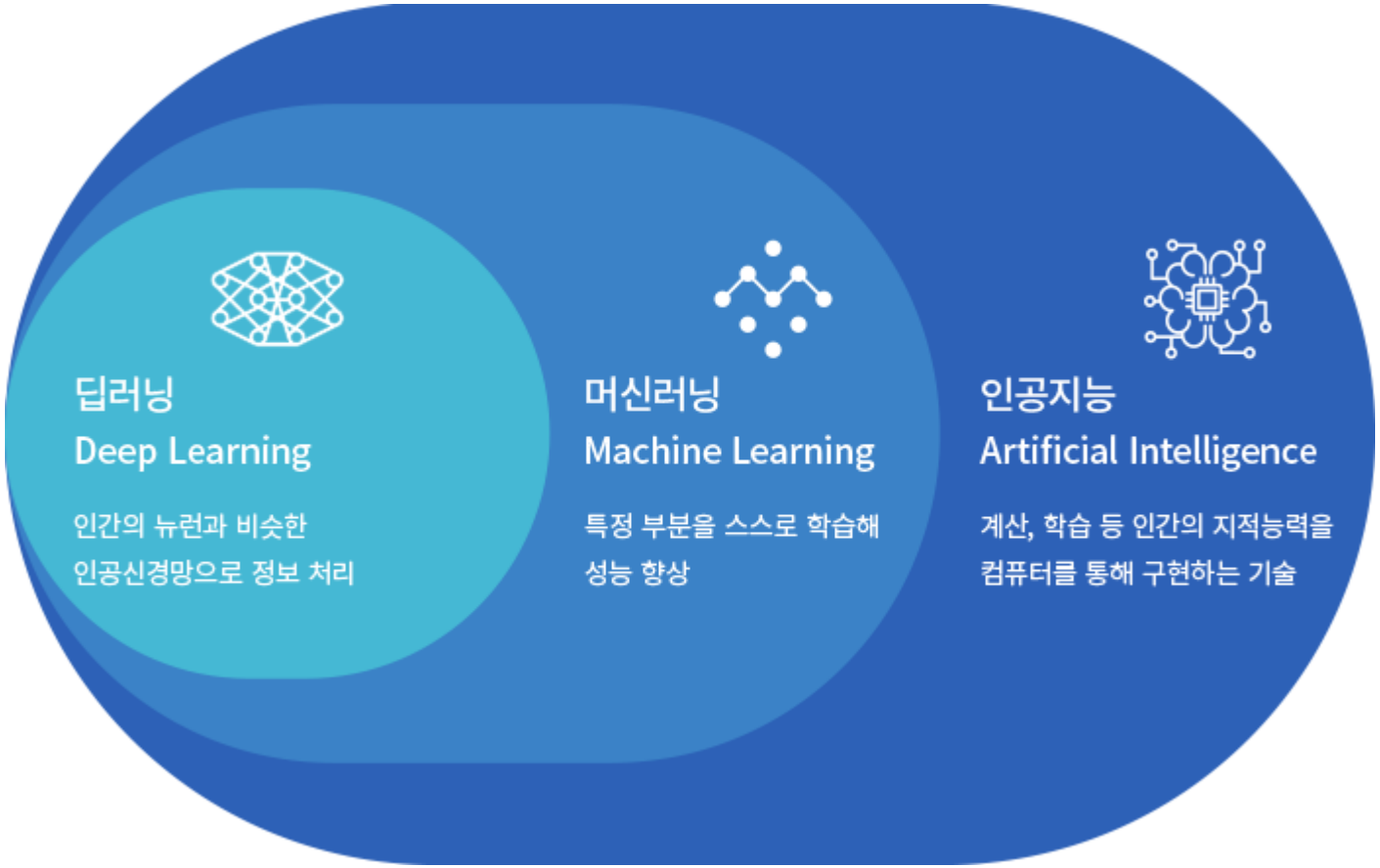
## 최상목 "한국형 챗 GPT 개발…3兆 AI스타트업 펀드 조성"

이승배 기자 · 2025. 2. 20. 11:37

27

崔대행, 제3 국가AI위원회 주재
"AI패권 경쟁 새 국면…국가역량 강화"
AI국가대표 선발해 LLM 프로젝트 가동
국가컴퓨팅센터 조속 구축…세제지원도

# 인공지능


딥러닝
Deep Learning
인간의 뉴런과 비슷한
인공신경망으로 정보 처리
머신러닝
Machine Learning
특정 부분을 스스로 학습해
성능 향상
인공지능
Artificial Intelligence
계산, 학습 등 인간의 지적능력을
컴퓨터를 통해 구현하는 기술

# 인공지능 - Trend

Key technologies
2035
Artificial intelligence
Human enhancement technologies
Sensors
Autonomous systems, incl. manned–unmanned teaming
Nanotechnology
Synthetic environments, virtual reality and augmented reality
Smart/ complex materials
Satellites and pseudo-satellites
Additive and advanced manufacturing
Directed energy weapons, electronic warfare, electronic countermeasures
Communications systems
Energy generation and storage

# 인공지능 - Trend

## 검색의 시대, 끝나다

AI쇼크

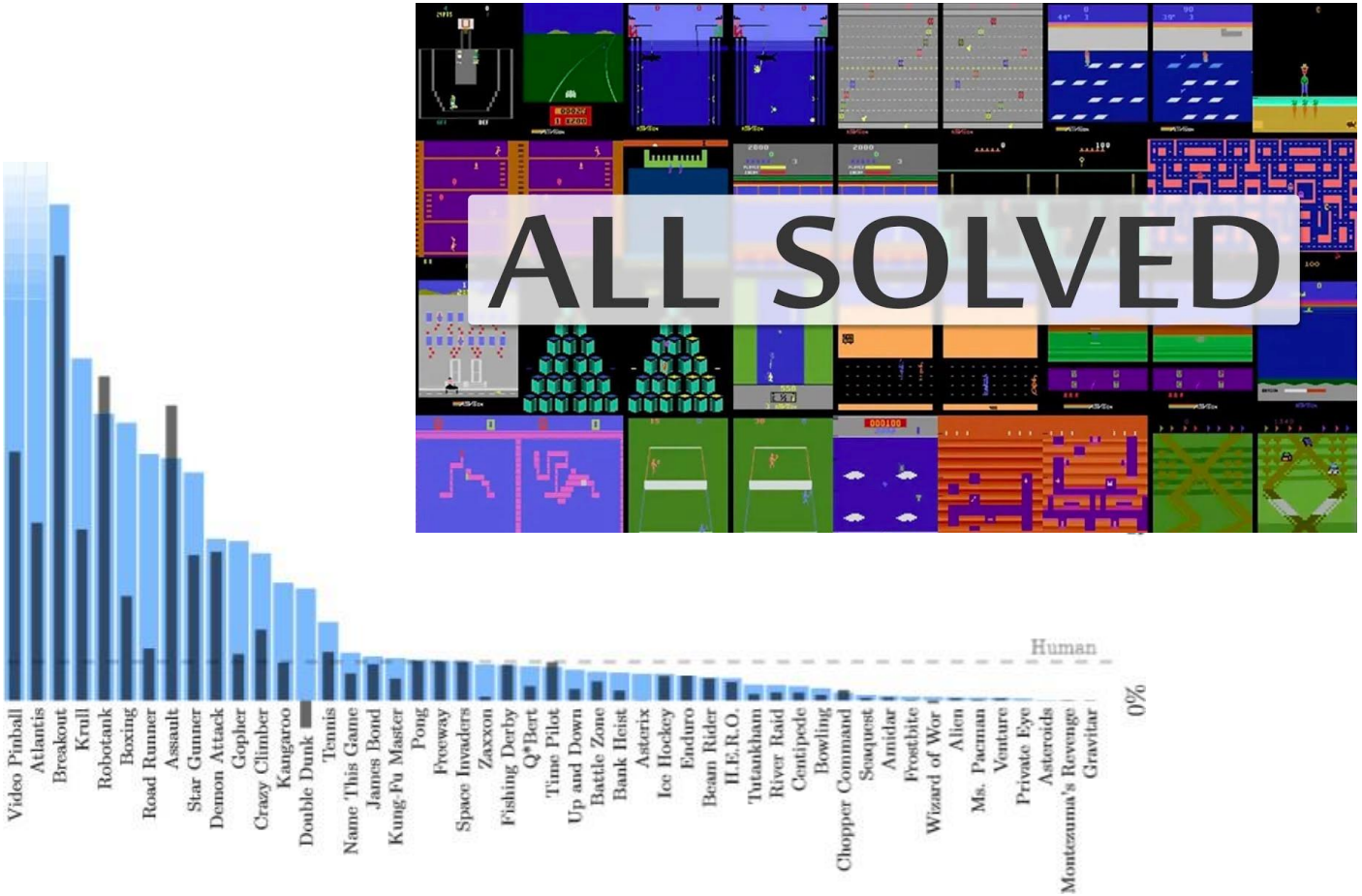
**정병태 교수** | 기사입력 2023/04/28 [17:53]

---

요즘도 인공지능(AI) 챗GPT가 전 세계적으로 엄청난 파장을 일으키고 있다.

글로벌 검색 시장은 구글이 83%를 넘는 압도적 점유율로 1위다. MS 빙이 2위에 있다. 그런데 구글이 위기의식을 느끼며 사활을 걸고 AI 경쟁력을 선보이고 있다. 구글 딥마인드(Google DeepMind)에서 새로운 대화형 대규모 언어 모델 바드, 람다(LaMDA)병기를 내놓았다. LaMDA는 Google에서 개발한 대화형 대규모 언어 모델이다(LaMDA: Language Model for Dialogue Applications). LaMDA는 어떤 종류의 사전 교육 없이도 거의 모든 것에 대해 정상적으로 대화할 수 있도록 되어 있다.

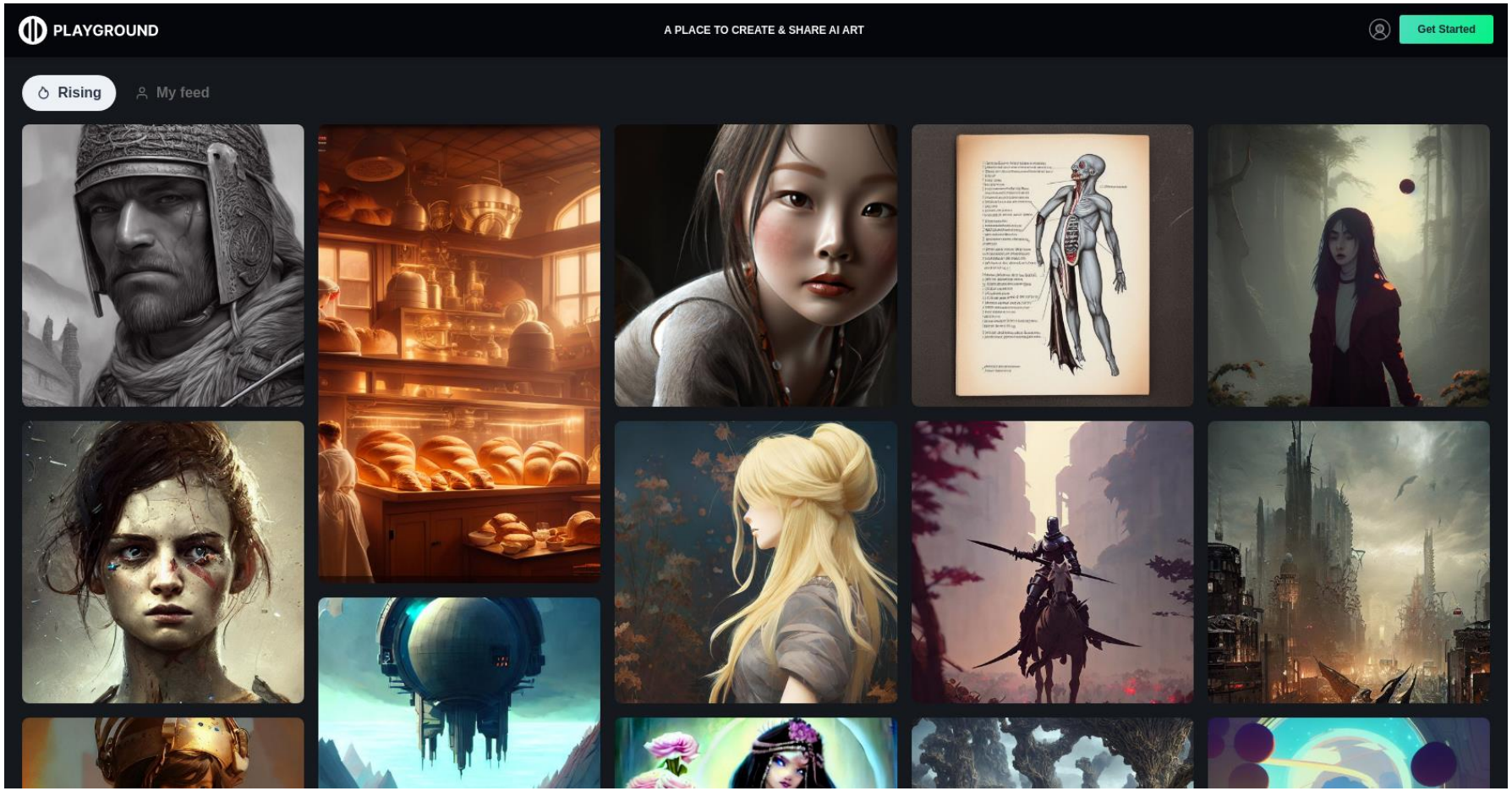
# 인공지능 - Trend




사람보다 게임을 잘 하는 Atari

Credit: Nvidia developer, youtube 'two minutes paper'

# 인공지능 - Trend

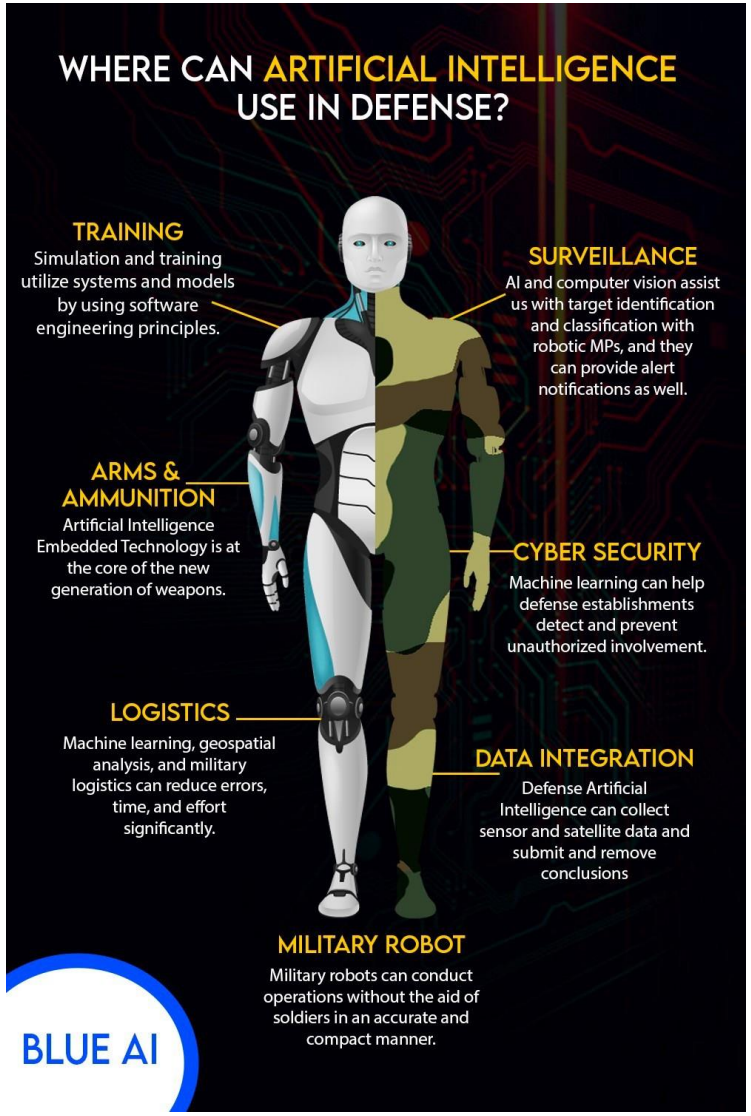

PLAYGROUND
A PLACE TO CREATE & SHARE AI ART
Get Started
Rising
My feed

# 인공지능 - Trend

계산대도 없는 무인마켓 'Amazon Go'

https://youtu.be/NrmMk1Myrxc

# 인공지능 - Trend


WHERE CAN ARTIFICIAL INTELLIGENCE USE IN DEFENSE?
TRAINING
Simulation and training utilize systems and models by using software engineering principles.
SURVEILLANCE
AI and computer vision assist us with target identification and classification with robotic MPs, and they can provide alert notifications as well.
ARMS & AMMUNITION
Artificial Intelligence Embedded Technology is at the core of the new generation of weapons.
CYBER SECURITY
Machine learning can help defense establishments detect and prevent unauthorized involvement.
LOGISTICS
Machine learning, geospatial analysis, and military logistics can reduce errors, time, and effort significantly.
DATA INTEGRATION
Defense Artificial Intelligence can collect sensor and satellite data and submit and remove conclusions
MILITARY ROBOT
Military robots can conduct operations without the aid of soldiers in an accurate and compact manner.
BLUE AI

YTN
news

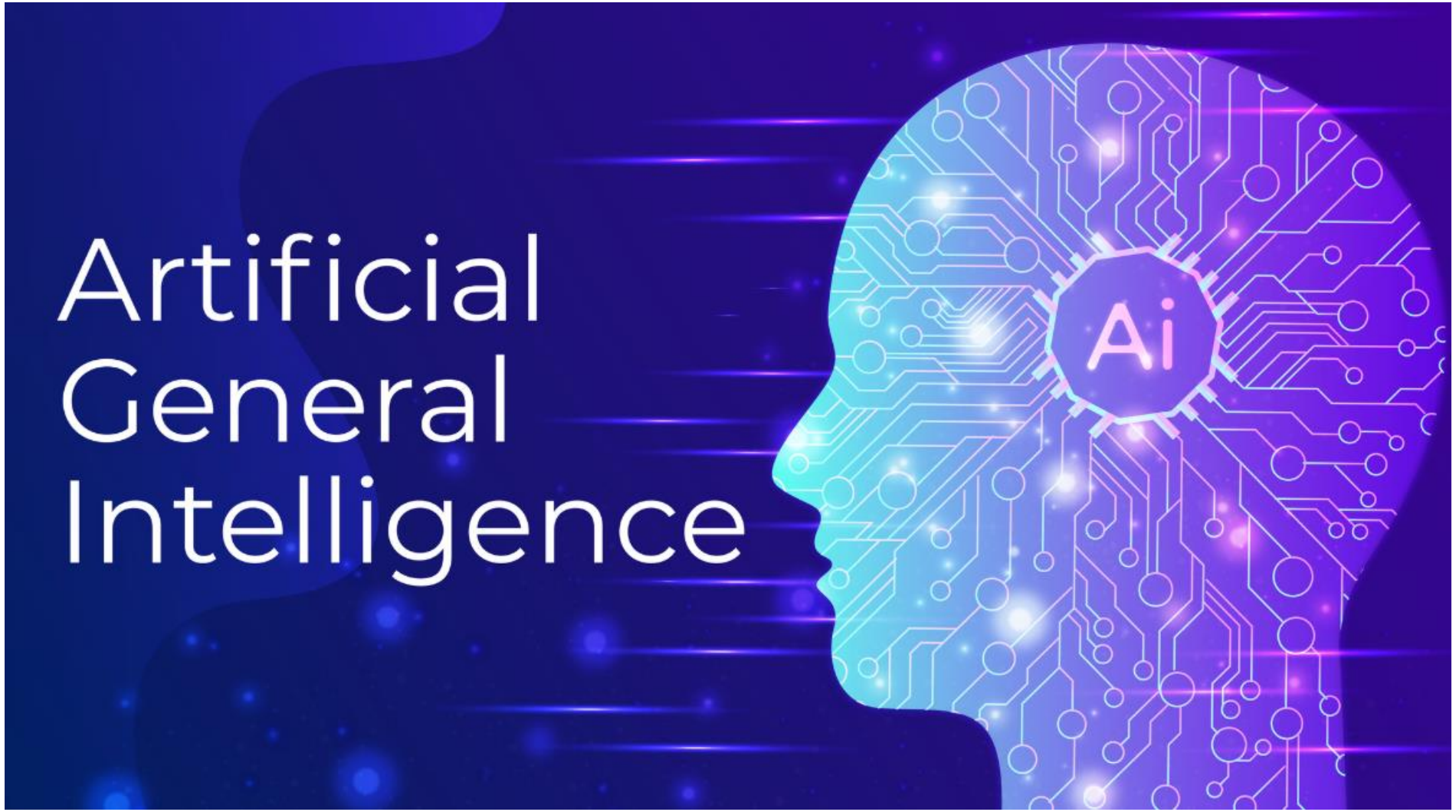
# 5. 인공지능 - Trend

✓ **AI Agent**

## 영화 – 이글아이 (2008)

https://m.blog.naver.com/jejus898/221439427034

# 5. 인공지능 - Trend

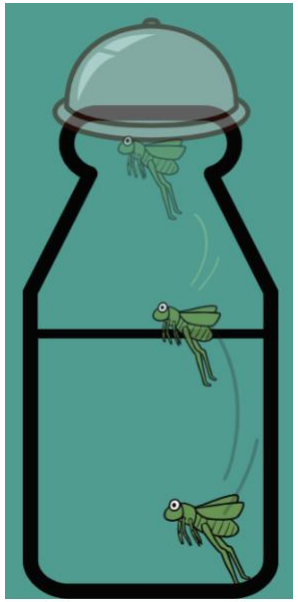


# 5. 인공지능 - Trend

deepseek

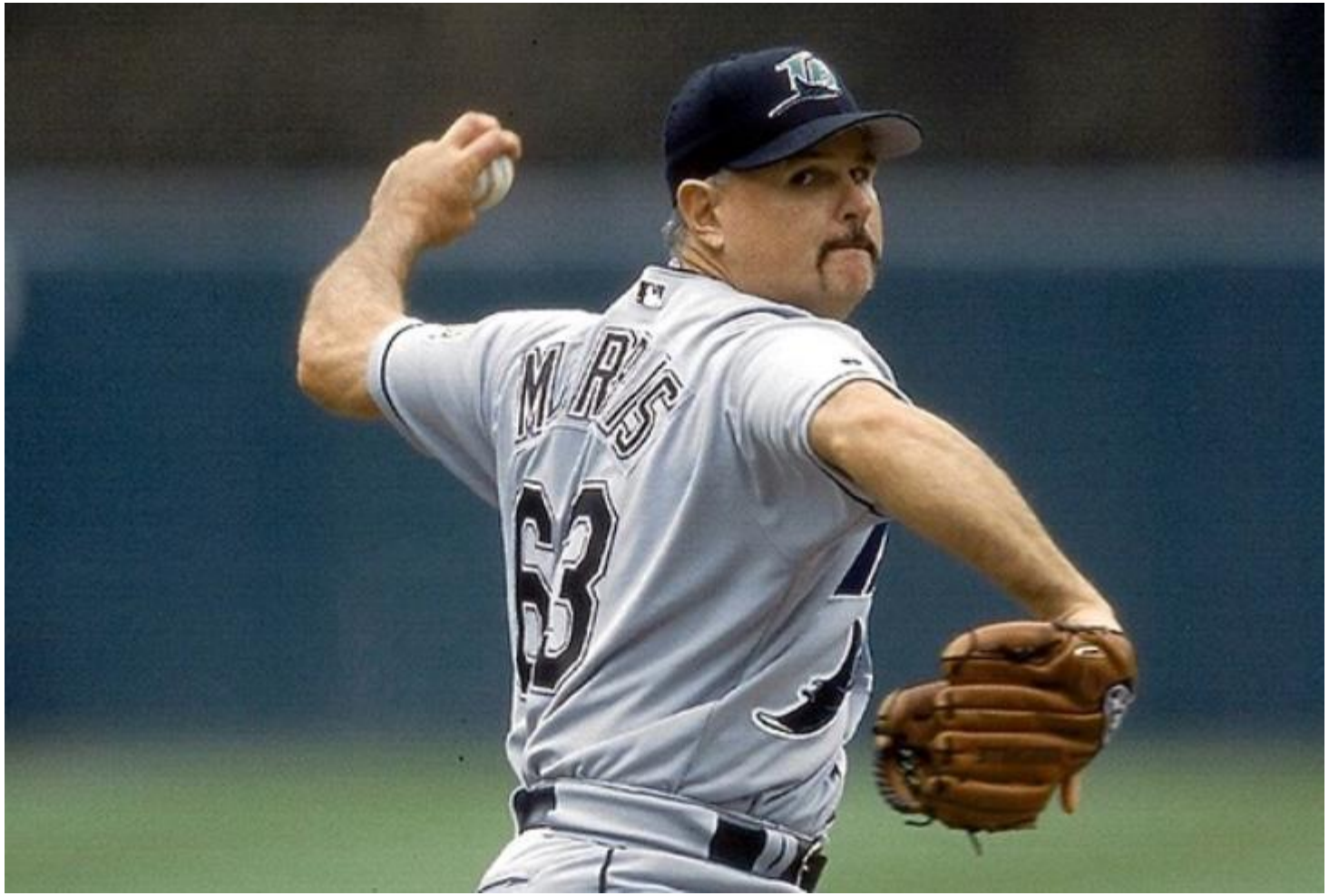
# 5. 인공지능 - Trend

# “자기 부하 한계”



- 최일구
- 다나카 고이치
- 이현승
- 정인덕
- .....

# Etc



# Etc

# 사람을 위한 기술

# THANK YOU