인공지능시대, 국가통계의 역할과 미래

서울대학교
통계학과
김용대

목차

1. 서론

빅데이터의 등장

YouTube

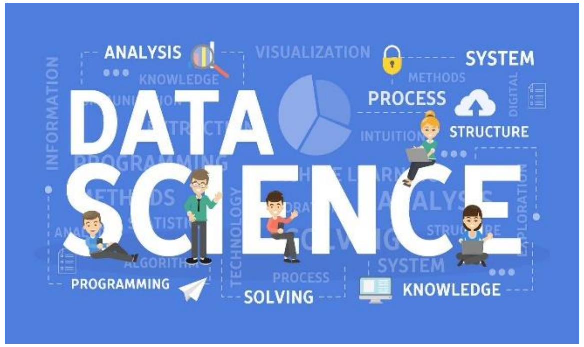
AT&T AdWorks

국가통계의 발전


통계청 보통씨가 알려주는
2022년 기준 경제통계 통합조사
2023년 6월 15일(목) ~ 7월 21일(금) 까지
인터넷조사
~ 7월 7일까지
대면조사
~ 7월 21일까지


organization
descriptive
scientific
Statistics
collection
interpretation
data
statistical
Inference
theory
experiments
mathematical
called
probability
parameters
mathematics

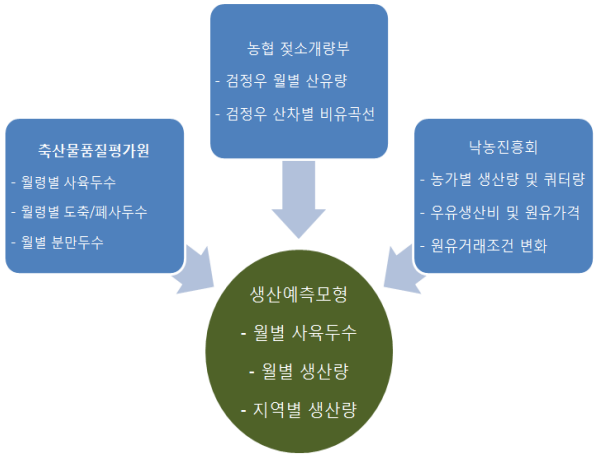

BIG
DATA


ANALYSIS
VISUALIZATION
SYSTEM
PROCESS
STRUCTURE
DATA
SCIENCE
PROGRAMMING
SOLVING
KNOWLEDGE

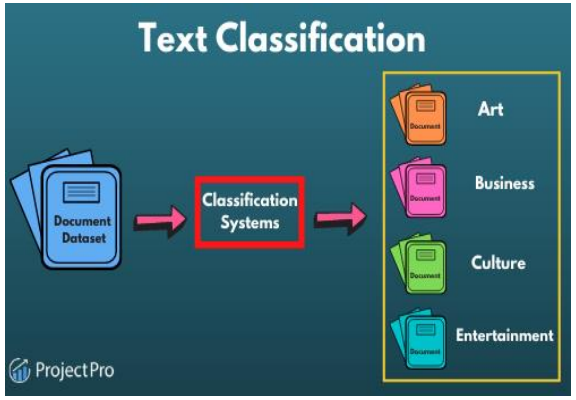
국가통계와 빅데이터

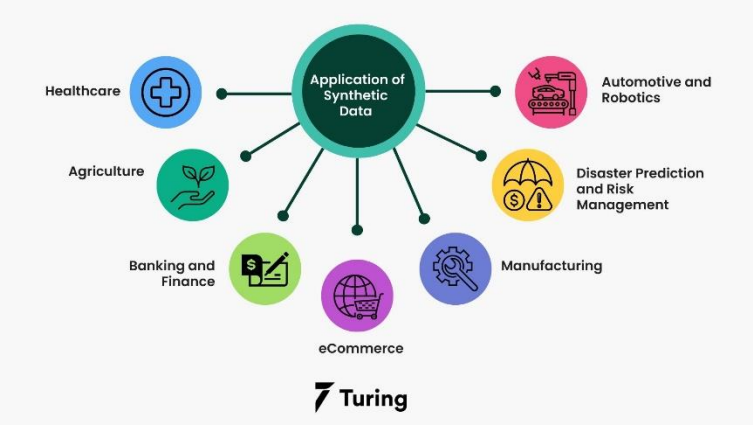
Nowcasting

내·외부의 **가축 전염병 관련 정보를 수집**, 이를 **분석**(가축전염병의 동향 및 국내영향도) 할 수 있는 Big Data 기반의 **가축 질병 발생 예측**과 **사전 차단 활동**에 활용



국가통계와 데이터사이언스





국가통계와 AI

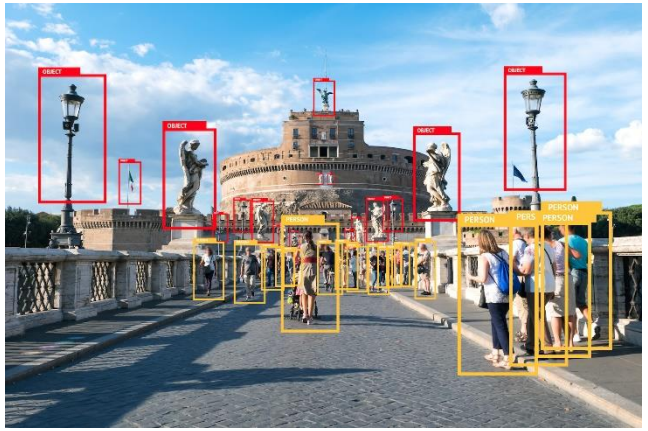
AI ≒ 데이터과학?

강연 내용

- AI와 데이터과학의 관계 설명
- AI를 국가통계에 적용하기 위한 방법론 예제
- 인공지능 시대, 국가통계의 발전을 위한 제언

2. AI, 데이터과학 그리고 국가통계

AI의 활약상



인간 지능의 자동화

AI와 데이터

- **AI를 위한 알고리즘을 데이터에서 찾음**
- **기계학습: 데이터에서 AI알고리즘을 찾는 방법론**
- **딥러닝: 기계학습의 하나의 방법론, AI에 매우 적합한 기계학습 알고리즘**
- **응용분야**
  - 이미지 인식
  - 음성인식
  - 언어인식
- **인간 지적 활동의 자동화**
- **데이터에 noise가 적음**

AI를 위한 데이터과학

인공지능 교과서에서는...

AI를 위한 데이터과학

실제는...

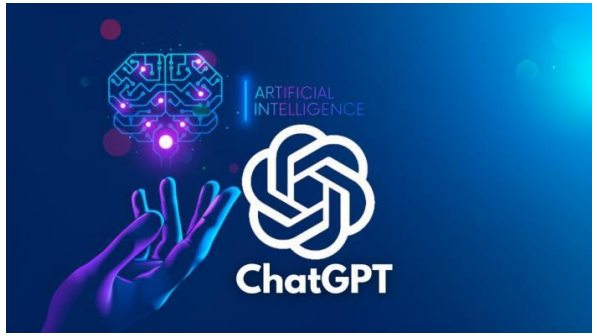
AI를 위한 데이터과학

데이터과학과 함께라면

AI를 위한 데이터과학

Attention Is All You Need

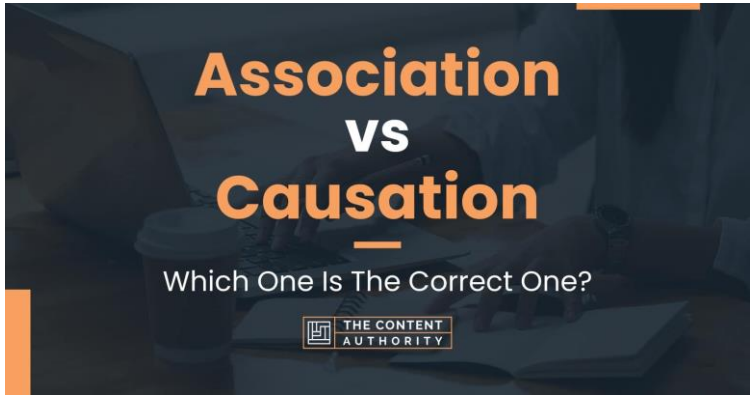
Ashish Vaswani* Google Brain avaswani@google.com
Noam Shazeer* Google Brain noam@google.com
Niki Parmar* Google Research nikip@google.com
Jakob Uszkoreit* Google Research usz@google.com
Llion Jones* Google Research llion@google.com
Aidan N. Gomez* † University of Toronto aidan@cs.toronto.edu
Łukasz Kaiser* Google Brain lukaszkaiser@google.com
Illia Polosukhin* ‡ illia.polosukhin@gmail.com



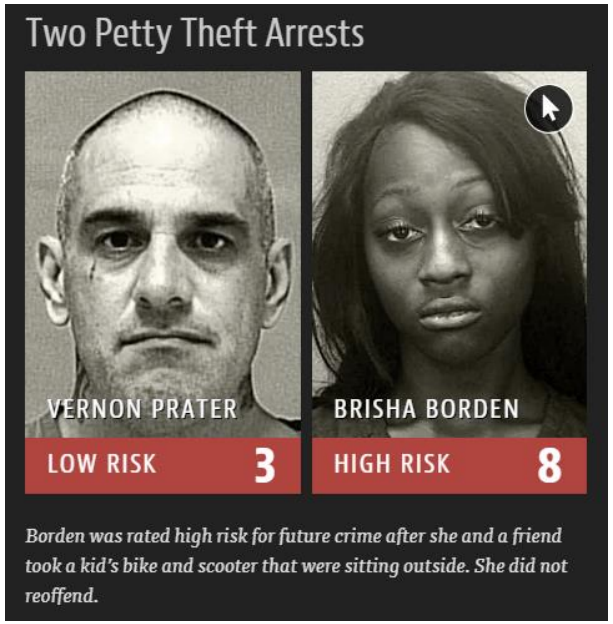
AI를 위한 데이터과학

What is Explainable AI (XAI) ?

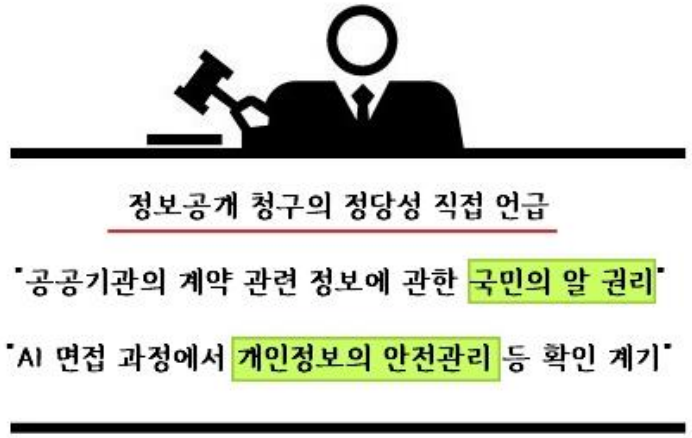
Source: https://miro.medium.com/max/1200/1*8ZPBdtMa-2k_EdXqQuB0LA.jpeg



AI를 위한 데이터과학

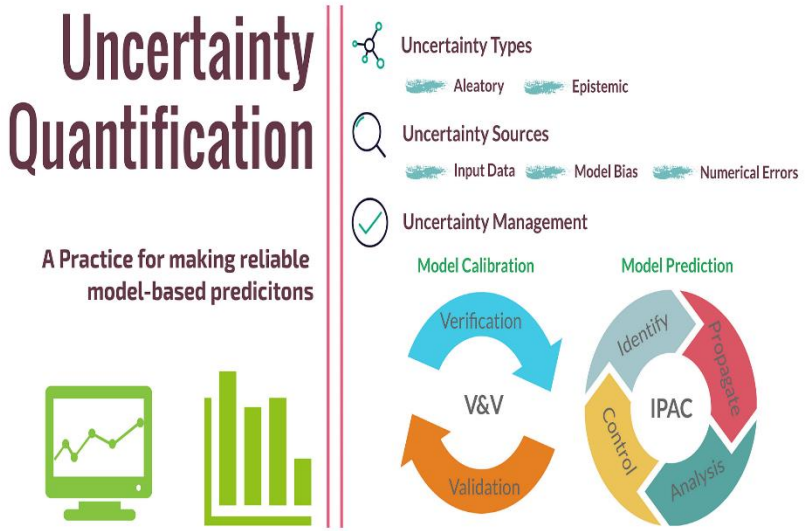

Two Petty Theft Arrests
VERNON PRATER
LOW RISK 3
BRISHA BORDEN
HIGH RISK 8
Borden was rated high risk for future crime after she and a friend took a kid's bike and scooter that were sitting outside. She did not reoffend.


AI FOR SCIENCE TRAINING SERIES
FAIR AI Models

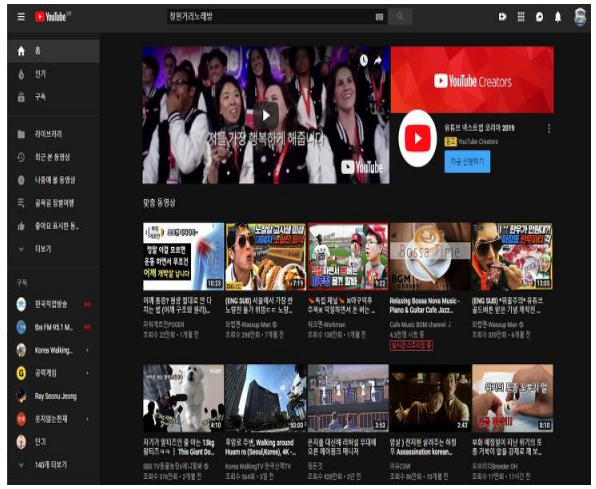

정보공개 청구의 정당성 직접 언급
"공공기관의 계약 관련 정보에 관한 국민의 알 권리"
"AI 면접 과정에서 개인정보의 안전관리 등 확인 계기"

AI를 위한 데이터과학


YTN NEWS

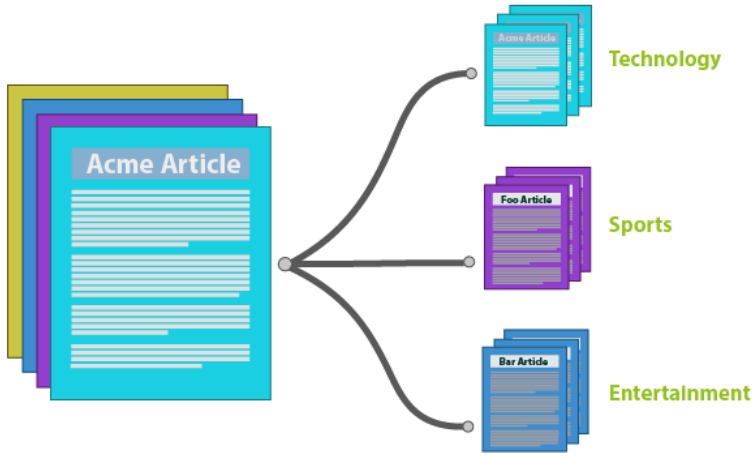

Uncertainty Quantification
A Practice for making reliable model-based predicitons
Uncertainty Types
Aleatory
Epistemic
Uncertainty Sources
Input Data
Model Bias
Numerical Errors
Uncertainty Management
Model Calibration
Model Prediction
Verification
V&V
Validation
Identify
Propagate
Control
Analysis
IPAC

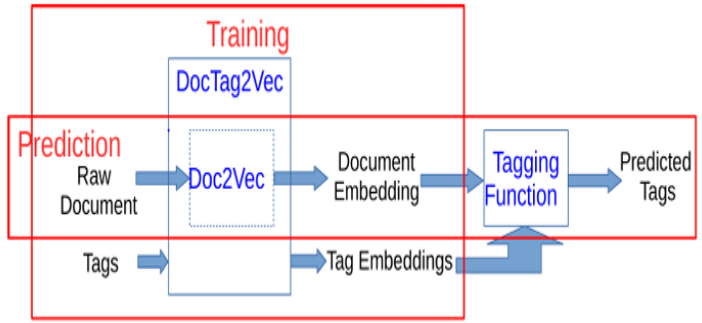
데이터과학의 활약상



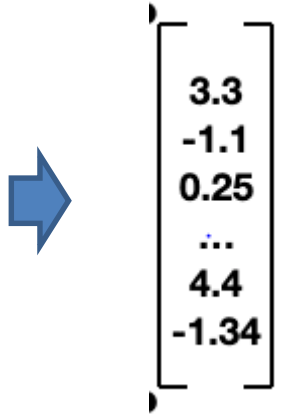
새로운 지식의 발견

데이터과학을 위한 AI



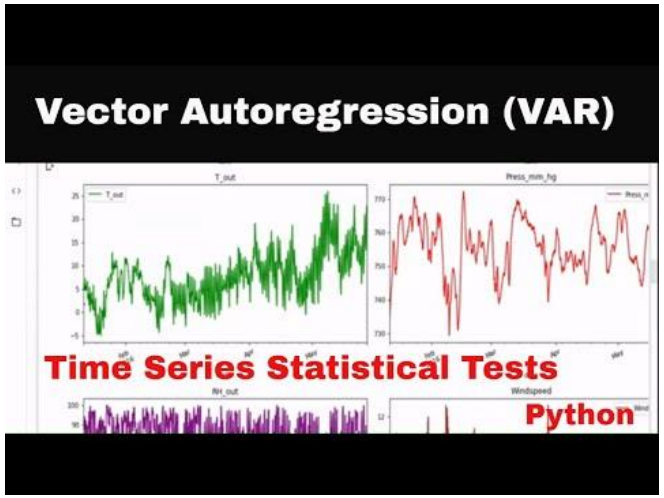


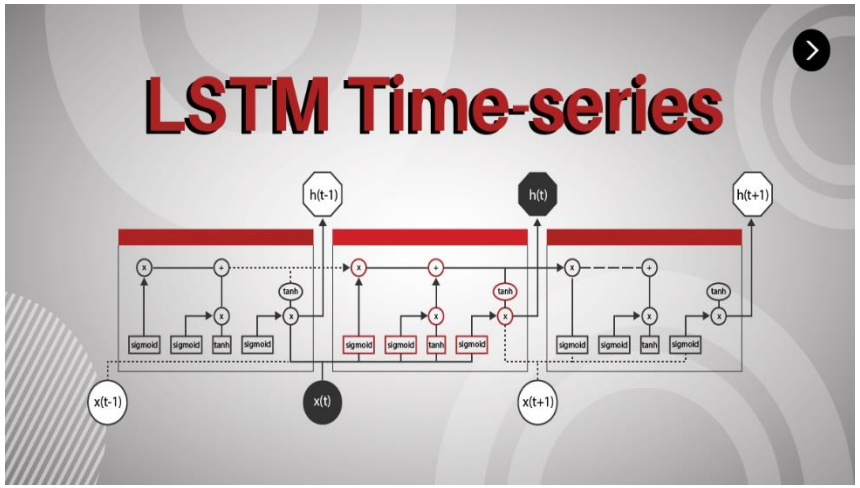
It was the best of times, it was the worst of times, it was the age of wisdom, it was the age of foolishness...



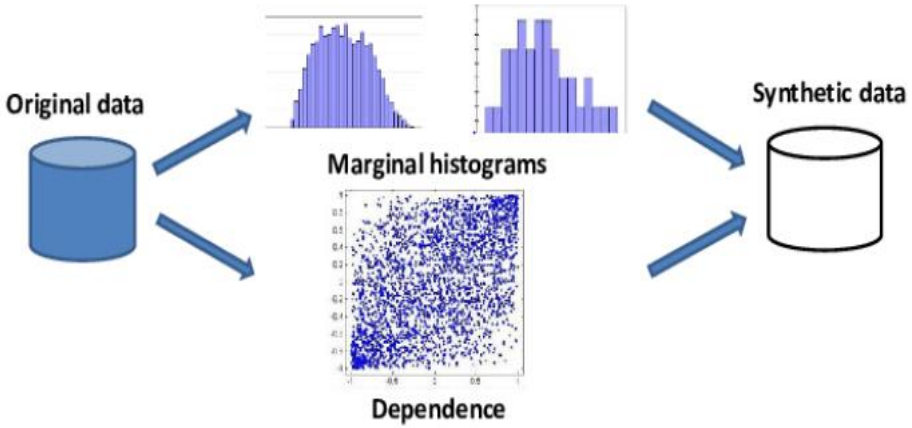
데이터과학을 위한 AI

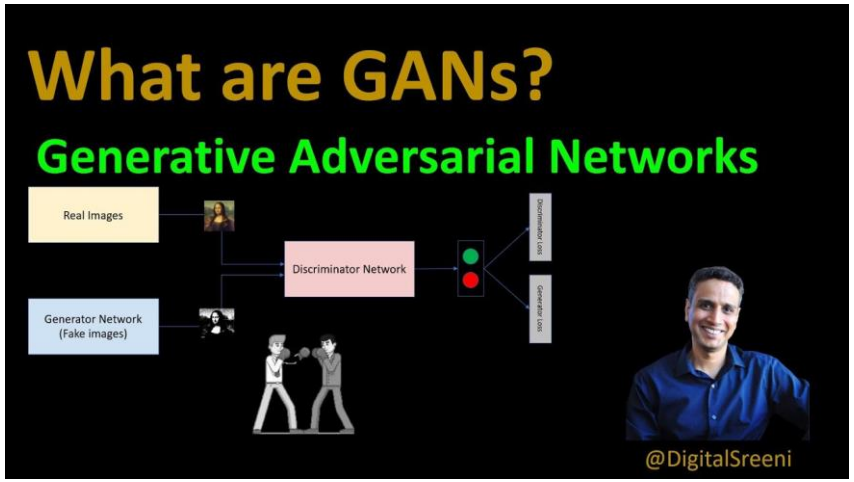

Time Series
Analysis
70.111
44.870
48.991
20.556
26.417
31.012
12.002
www.educba.com


Vector Autoregression (VAR)
Time Series Statistical Tests
Python


LSTM Time-series
h(t-1)
h(t)
h(t+1)
x(t-1)
x(t)
x(t+1)
sigmoid
tanh

데이터과학을 위한 AI


Original data
Marginal histograms
Synthetic data
Dependence

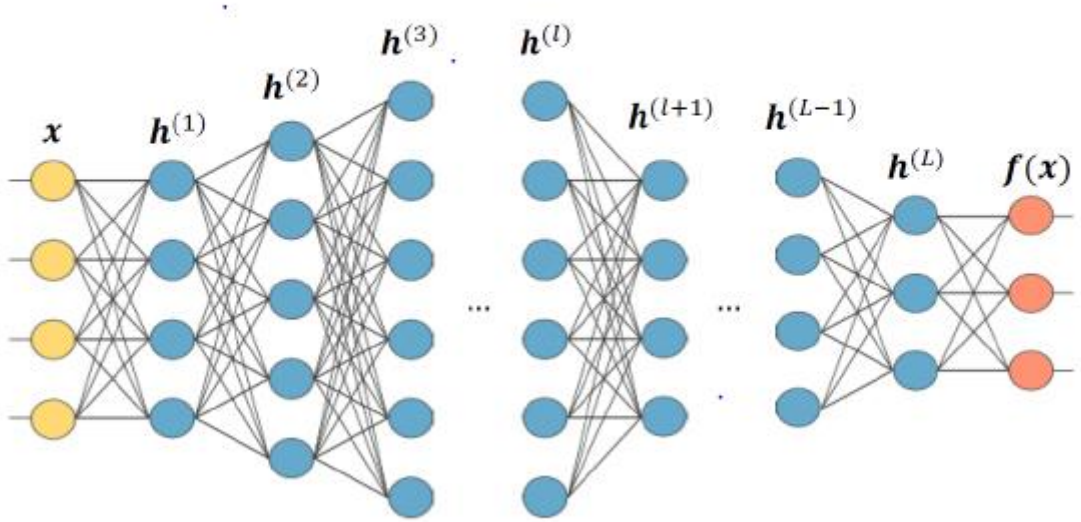

What are GANs?
Generative Adversarial Networks
Real Images
Generator Network
(Fake images)
Discriminator Network
@DigitalSreeni

국가통계를 위한 AI

- 인공지능 기술을 이용한 국가통계를 위한 데이터 획득, 관리, 분석의 효율화
- 적용가능 분야
  - 자동코드분류: 산업재해, 개인사업
  - 조사: 응답률 예측, 결측치 대체
  - 품질관리: 이상치 탐색, 데이터 정합성 확인
  - 비정형데이터 관리 및 분석: 문서데이터 분류, 인공위성 데이터 분석
  - 예측: 빅데이터 기반 Nowcasting, 전염병 확산 예측
  - 등등. ....

3. AI방법론의 국가통계 적용 예제: 이상치 탐색

딥뉴럴네트워크



딥뉴럴네트워크

- $\mathbf{z}^{(l+1)} = A_l \mathbf{h}^{(l)} + \mathbf{b}_l$
- $\mathbf{h}^{(l+1)} = \sigma(\mathbf{z}^{(l+1)})$ for $l = 0, \ldots, L-1$
- $\Pr(Y = j|\mathbf{x}) = f_j(\mathbf{x}) \propto \exp(-\mathbf{h}^{(L)}(\mathbf{x})^\top \gamma_j)$.

Memorization Effect

- 딥뉴럴네트워크의 특징 중 하나는 모형이 매우 유연하여 데이터를 완벽하게 분류할 수 있슴 (easily overfit)
- 딥뉴럴네트워크가 데이터를 과적합(overfit)할 때, 학습이 진행되면서 정상데이터를 먼저 학습하고 이상치를 나중에 학습하는 현상이 있고, 이러한 현상을 **Memorization effect (ME)** 라고 함.
- Label noise 문제 (분류문제에서 label이 잘 못 부여된 데이터 탐지)에 널리 사용됨

Label Noise 문제

- 데이터: $(\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_n, y_n)$ 이고 $\mathbf{x}_i \in \mathbb{R}^p, y_i \in \{-1, 1\}$.
- Label noise 문제
  - y_i^{gt}를 i번째 관측치의 실제 속하는 그룹의 라벨
  - 몇몇 관측치에서 관측라벨 y_i이 실제라벨 y_i^{gt}와 다른 경우
  - 분석의 목적은 라벨이 잘못 붙여진 데이터 $\{i : y_i \neq y_i^{gt}\}$를 찾는 것이다.
  - 이 문제의 어려운 점은 실제라벨 y_i^{gt}이 전혀 관측되지 않는다는 것이다.

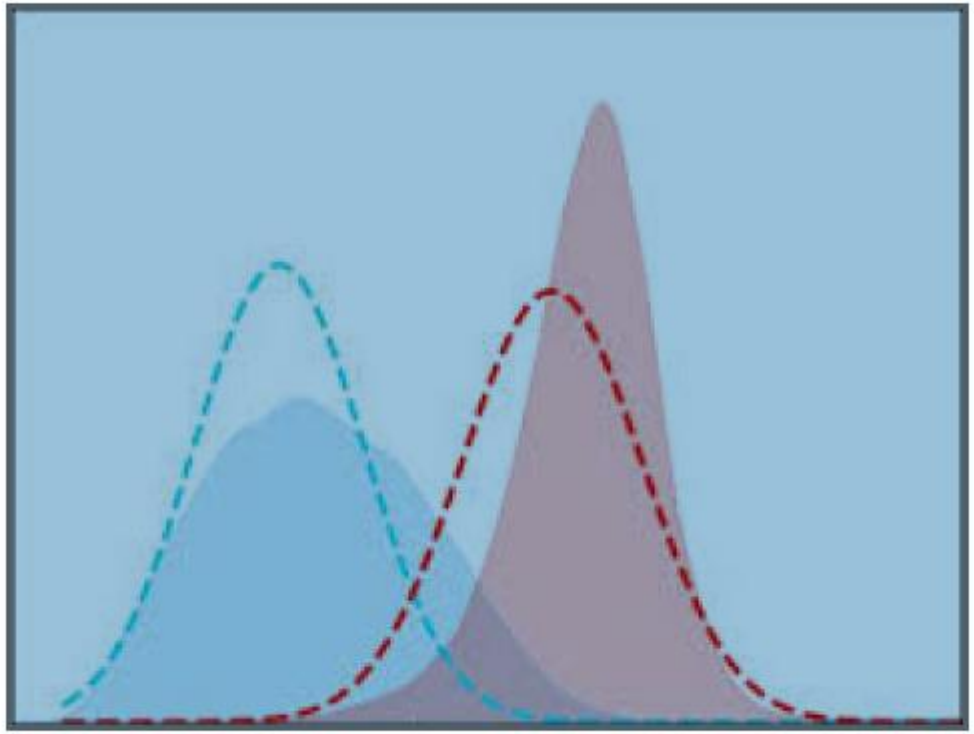
Label Noise 문제가 생기는 이유

- 보통 라벨은 사람이 붙이는데, 사림의 실수로
- 최근에는 인공지능을 이용한 라벨링이 널리 사용되는데 (예: 개인사업자 산업분류), 인공지능 알고리즘이 부정확해서
- Label noise알고리즘을 적용해서 탐지된 데이터에 대해서, 전문가가 한번 더 확인하므로써, 데이터의 품질 향상 가능

Label Noise 탐지 방법

- 먼저, 딥뉴럴네트워크를 데이터에 살짝 적합한다.
- 여기서, '살짝'이란 심딥뉴럴네트워크를 학습하는 알고리즘을 10번 미만으로 돌린다 (참고: 보통은 수천번 돌림).
- 살짝 학습된 딥뉴럴네트워크가 잘 못 맞추는 데이터를 골라내서 noise label데이터라고 판단한다.

Label Noise 탐지 방법



이상치 탐지 방법

- 딥뉴럴네트워크의 Memorization effect을 성질을 이용해서 이상치 탐색도 가능함.
- 이상치 탐색을 하는 일반적인 프로세스
  - 데이터의 분포를 추정한다 (정규분포가정, 비모수적 방법)
  - 데이터에 이상치가 있기 때문에, 보통 robust한 방법으로 분포를 추정한다.
  - 추정된 분포에서 많이 벗어나는 데이터 (추정된 분포에서 확률이 작은 데이터)를 이상치로 판단한다.

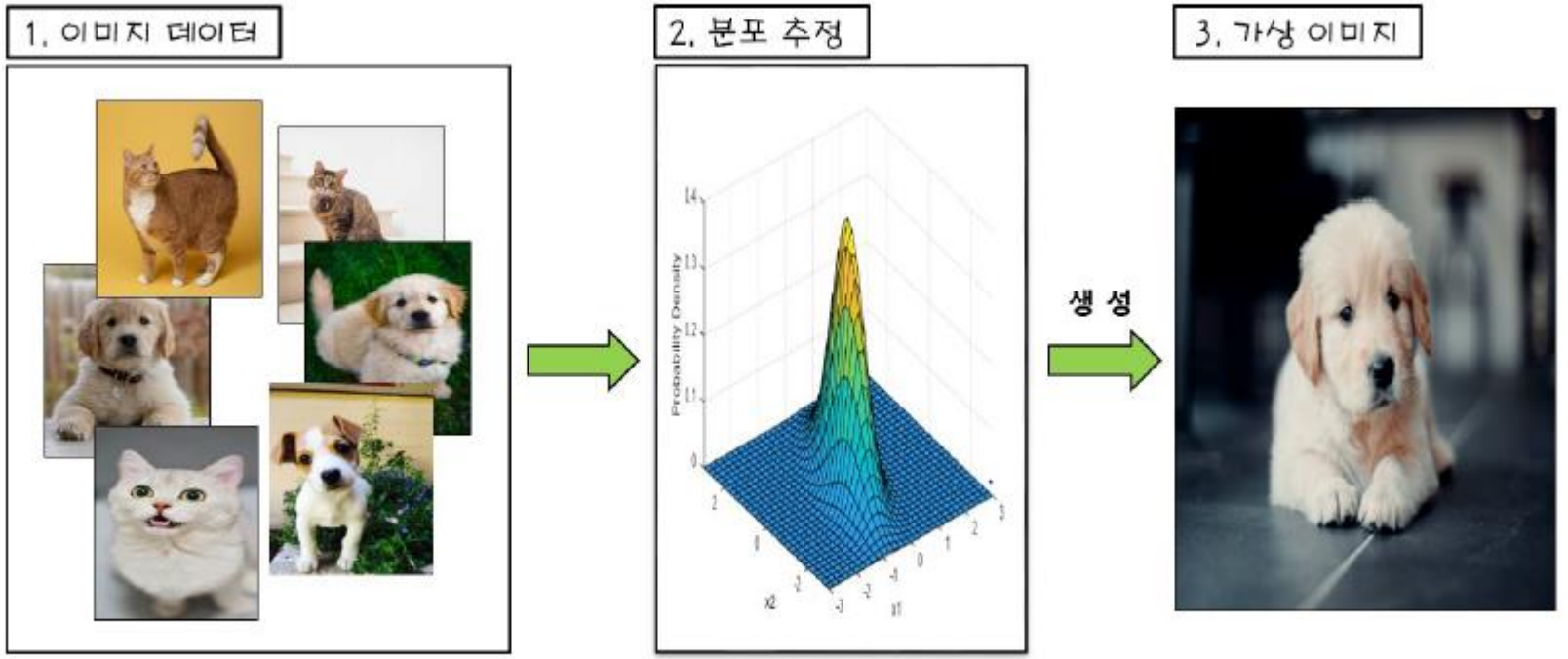
이상치 탐지 방법

- 일반적인 이상치 탐지 프로세스의 어려움
  - Robust하게 분포를 추정하는 것이 기술적으로 매우 어려움.
  - 정규분포를 가정하면 상대적으로 쉬우나, 비모수적 robust분포 추정은 쉽지 않음.
  - 특별히, 데이터의 다중공선성이 강한 경우 일반적인 방법들의 효율이 급격히 떨어짐.
- 분포를 추정하는 딥뉴럴네트워크의 Memorization effect를 이용하면 매우 효율적으로 이상치 탐색이 가능함!

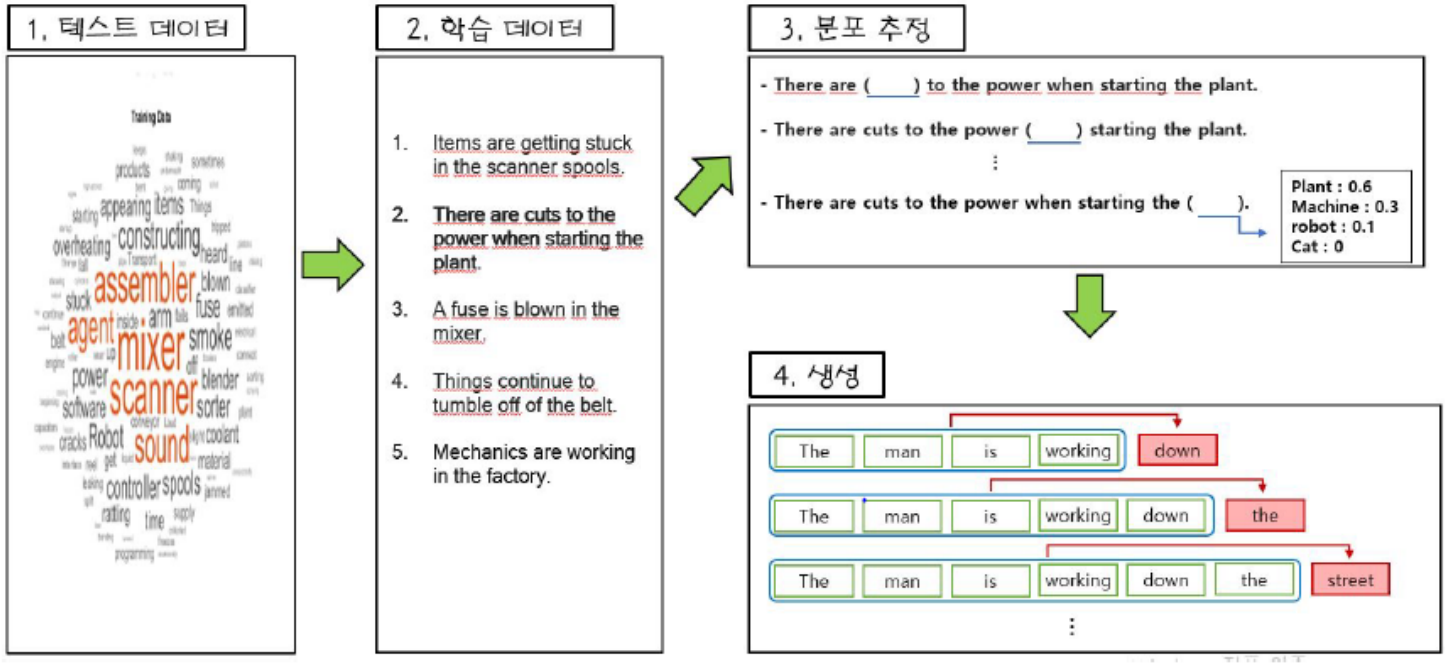
분포추정 딥뉴럴네트워크

- 딥생성모형 (Deep generative model): 딥뉴럴네트워크로 데이터의 분포를 추정하는 알고리즘
- 최근에 엄청난 유명세를 타고 있는 ChatGPT도 분포를 추정하는 딥생성모형 중 하나임.
- 이미지생성: 주어진 이미지데이터의 분포를 추정한 후, 추정된 분포에서 이미지 생성
- Chat GPT: 주어진 문서데이터에서, 다음에 나오 단어의 분포 (확률)을 추정하고, 이를 이용하여 문장 생성.

분포추정 딥뉴럴네트워크



분포추정 딥뉴럴네트워크



딥생성모형을 이용한 이상치 탐지

- 주어진 데이터를 딥생성모형으로 살짝 적합한다.
- 여기서, '살짝'이란 딥생성모형을 학습하는 알고리즘을 10번 미만으로 돌린다 (참고: 보통은 수천번 돌림).
- 살짝 학습된 딥생성모형이 잘 생성하지 못하는 데이터를 골라내서 이상치로 판단한다.
- 예제
  - 정상: '아버지가 방에 들어가신다.'
  - 이상치: '아버지 가방에 들어가신다.' ('아버지'는 '가방에' 거의 들어기시지 않음. 즉 이런 문장은 거의 생성이 안됨).

실험결과

Table 1: Averaged AUC and AP scores over 30 tabular datasets.

Method	OCSVM	LoF	IF	COPOD	ECOD	DeepSVDD	RSRAE	ODIM
AUC	0.706	0.649	0.781	0.749	0.749	0.643	0.656	**0.790**
AP	0.309	0.228	0.381	0.335	0.343	0.196	0.277	**0.398**

Table 2: Results of AUC scores on image and text data.

Method	OCSVM	LoF	IF	COPOD	ECOD	DeepSVDD	RSRAE	ODIM
MNIST	0.847	0.703	0.819	0.774	0.767	0.753	**0.902**	0.836
FMNIST	0.874	0.497	0.909	0.867	0.843	0.808	0.844	**0.909**
CIFAR10	0.659	0.675	0.878	0.902	0.884	0.617	0.866	**0.921**
MNIST-C	0.726	0.567	0.717	0.714	0.720	0.540	0.710	**0.740**
MVTec-AD	0.726	0.851	0.832	0.805	0.816	0.639	0.801	**0.900**
WM-811K	0.731	0.327	0.680	0.704	0.695	0.505	0.672	**0.747**
20News	0.649	0.727	0.645	0.646	0.645	0.513	0.636	**0.702**
Agnews	0.658	0.764	0.681	0.686	0.672	0.522	0.649	**0.804**
Idmb	0.496	**0.524**	0.517	0.499	0.494	0.485	0.491	0.519
Yelp	0.588	0.661	0.611	0.607	0.580	0.505	0.589	**0.669**

4. 결론 및 제언

결론

- DNN은 단순히 기존 방법론의 대체재가 아니라, 새로운 도구이다!

- AI방법론은 국가통계의 질적 향상에 도움이 된다!

제언

- 그러나, 모든 AI방법론이 국가통계에 도움이 되는 것은 아니다.
- 예: DNN은 noise가 많은 데이터에서는 예측력이 안 좋음.
- 예: DNN은 robust하지 않음.

- 예: GAN은 테이블데이터에서는 성능이 안좋음 (Table GAN 개발)

제언

- 국가통계를 위한 AI방법론에 대한 교육 및 연구가 필요하다!
- 그리고, 이를 실무에 적용하기 위한 조직도 따로 필요하다!


Thank
You
Merci
Danke
谢
谢
Asante
Kiitos
köszönöm
Maake
Shukria
Dhanyavadagalu
Vinaka
Kam Sah Hammida
شكرا
Manana
Dankon
감사합니다
Dank Je
Mauruuru
Biyan
Matondo
Blagodaram
Ngiyabonga
Dziekuje
Chokrane
Arigato
Grazie
Tack
Juspaxar
Gracias
Mochchakkeram
நன்றி
धन्यवाद
cảm ơn bạn
Tingki
Ua Tsaug Rau Koj
Gratias Tibi
Děkuji
Nirringrazzjak
Obrigado
ありがとう
Suksama
Hvala
Welalin
Djiere Dieuf
Misaotra
Go Raibh Maith Agat
Eskerrik Asko
Grazas
Salamat
ขอขอบคุณคุณ
Kia Ora
Kop Khun Khap
Paldies
Terima Kasih
Diolch i Chi
Taiku
Tack
Matur Nuwun
Rahmat
Di Ou Mèsi
Dankscheen
спасибо
Bedankt
Dakujem
Najis Tuke