Hewlett Packard Enterprise | NVIDIA | DX 동국시스템즈

HPE - NVIDIA - 동국시스템즈

AI Solution Day

2024년 4월 9일(화), 09:00~13:30

그랜드 인터컨티넨탈 파르나스, 로즈(5F)

Hewlett Packard Enterprise | NVIDIA | DX 동국시스템즈

HPE - NVIDIA - 동국시스템즈
AI Solution Day

How to Find Quality Data in AI Landscape

급변하는 AI 시장에서 길 찾기

Testworks 김성현 팀장

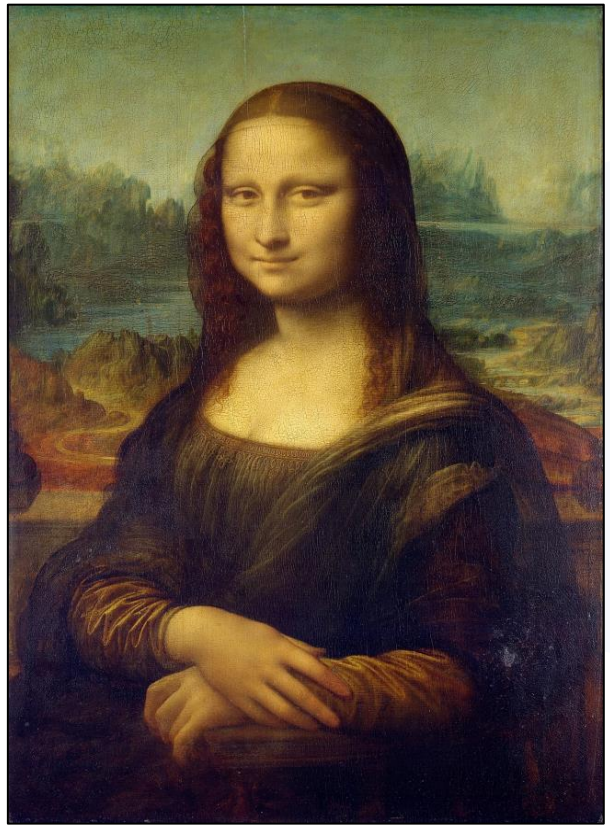
Index

1. Introduction

1

What happened to paintings?

Mona Lisa (1503-1506)



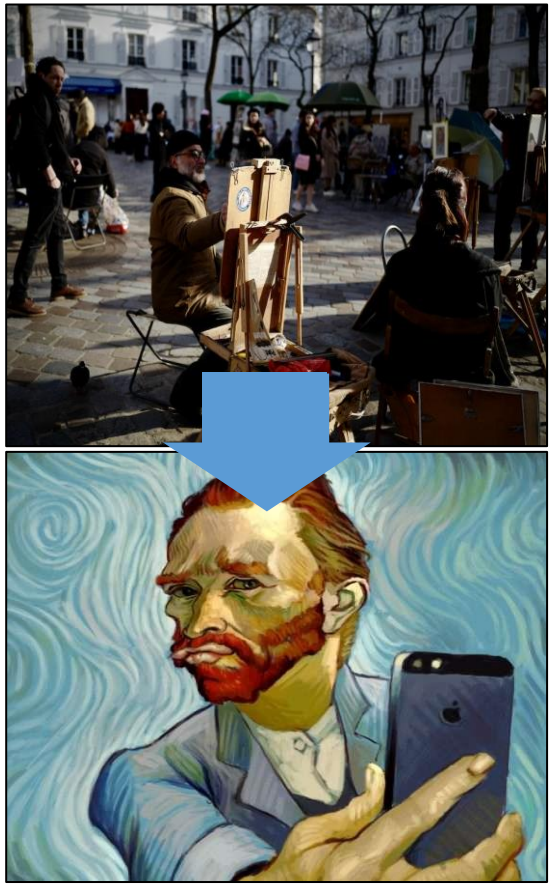
Camera
(~1826)

The Weeping Woman (1937)

Image Source: https://en.wikipedia.org/wiki/Mona_Lisa, https://www.masterclass.com/articles/when-was-the-camera-invented, https://en.wikipedia.org/wiki/The_Weeping_Woman

2

The impact of the camera

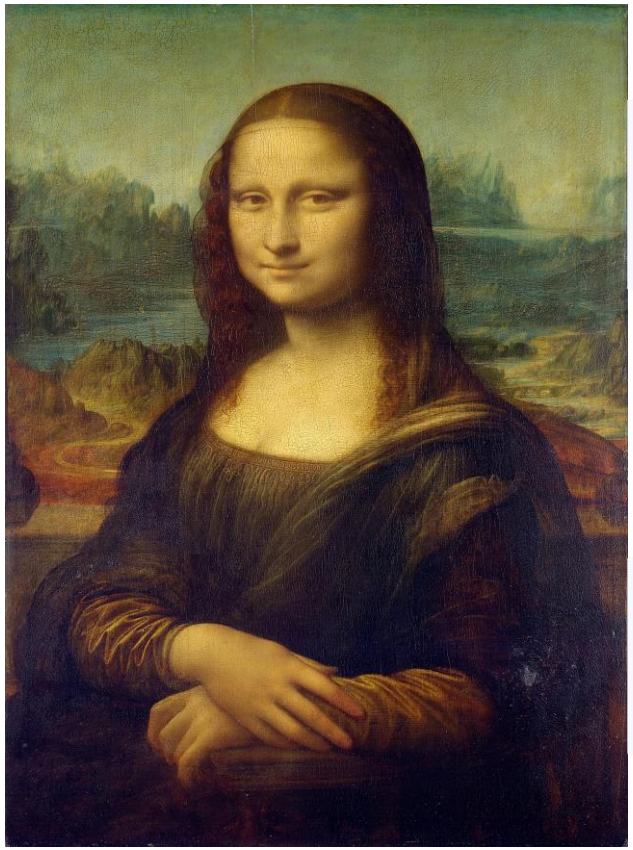


Picture Realism → Photo Realism

- **Democratization of visual art:** Before the camera, only the wealthy could afford portraits.
- **Job displacement:** The invention of the camera made the job of painters disappear or change their job descriptions.
- **Privacy Concerns:** The camera's ability to capture images led to concerns about privacy. People could be photographed without their consent, leading to issues of invasion of privacy, especially in public spaces.
- **Misrepresentation and Manipulation:** Photos could be altered or staged to convey false information or narratives.

Image Source: https://unsplash.com/photos/O_vFB1K0ttk, https://www.artculturefestival.in/evolution-self-portraiture

AI is the new sheriff in town!



Camera

1800 ~

Picture Realism → Photo Realism

AI

2020 ~

Image Source: https://en.wikipedia.org/wiki/Mona_Lisa , https://en.wikipedia.org/wiki/The_Weeping_Woman

Mona Lisa deepfakes developed the Samsung AI Center in Moscow

Painting went multi-modal too

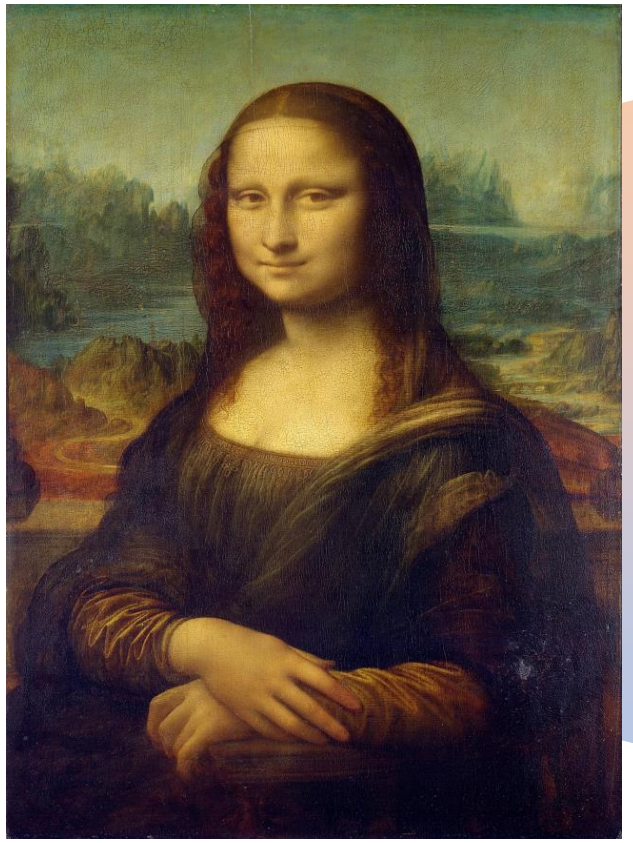


Image Source: https://en.wikipedia.org/wiki/Mona_Lisa

Camera

1826

+

Phonograph

1877

What's Next?

Film success: Synchronization

- **For multimedia, timing is everything!**
  - Why are there holes around 35 mm films?
  - "Film perforations" to synchronize and play the film at a constant speed

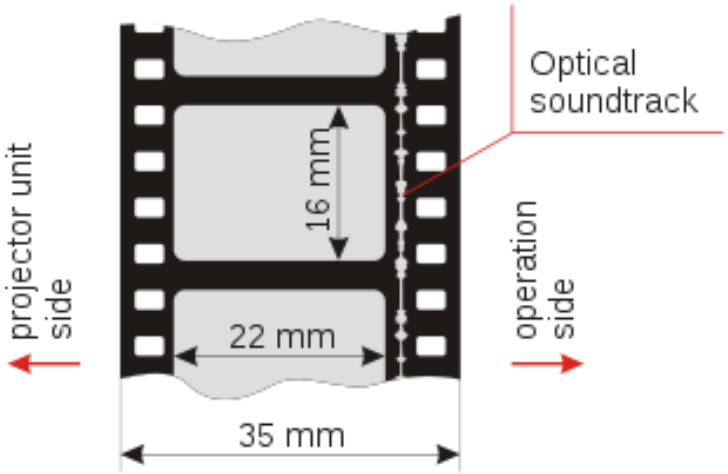
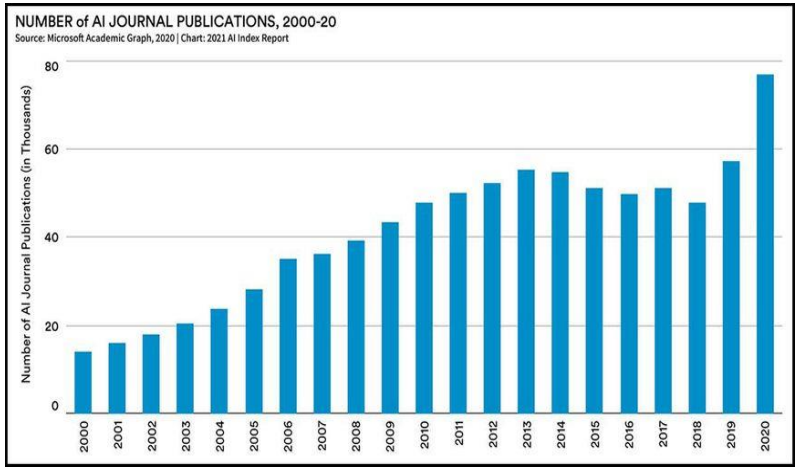


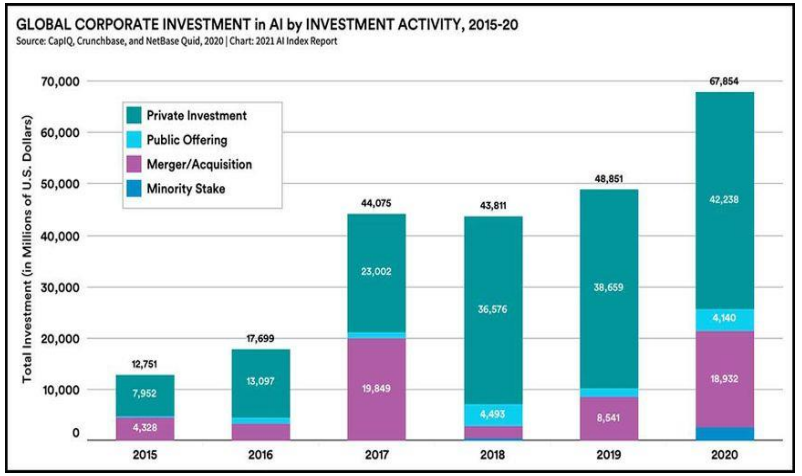
Image Source: https://en.wikipedia.org/wiki/35_mm_movie_film

6

AI's impact to society and industry

- "**Every company will become an AI company**." (2020)
  - Arvind Krishna (IBM CEO)
  - Every industry will be transformed
  - Look what happened to Amazon, Starbucks, Boeing, etc.
- **Search:** auto-correct, search rankings, understanding search queries, voice search, image search
- **Machine translation:** many to many language translations are possible and the list is still growing
- **Autonomous driving:** Tesla, Waymo, etc.
- **Financial:** smart banking, institutional investors (trading bots)
- Smart farm, smart factory, smart city, etc.

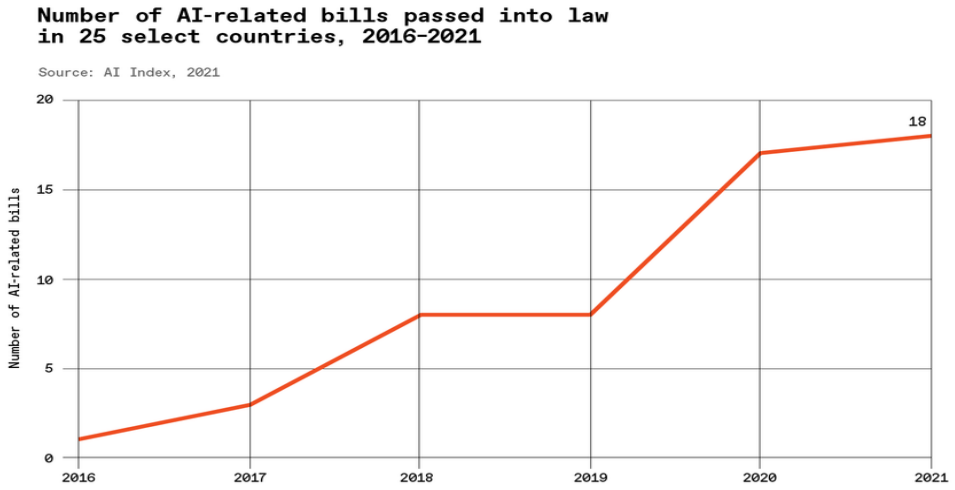




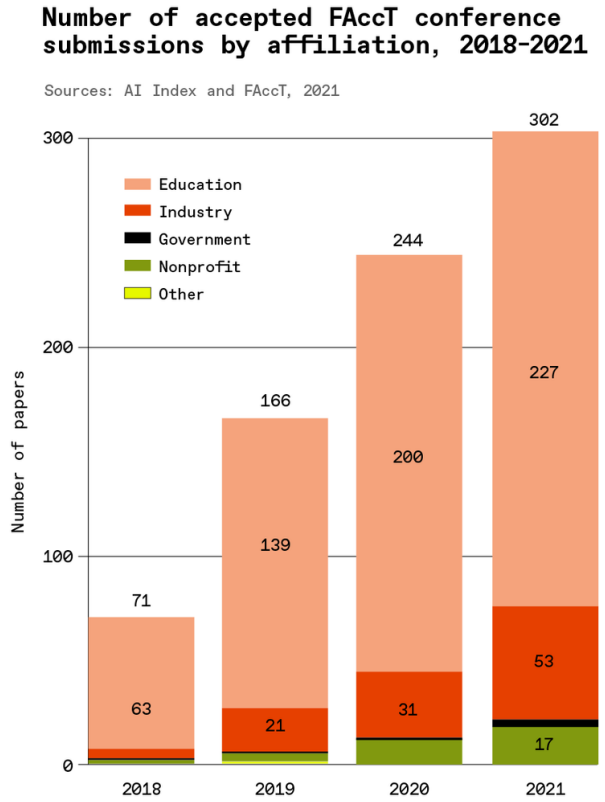
https://spectrum.ieee.org/the-state-of-ai-in-15-graphs

More concerns about AI ethics

- More interests in ACM conference on Fairness, Accountability, and Transparency (FAccT) and ethics-related workshops at NeurIPS.
- Law makers are paying attention too



https://spectrum.ieee.org/artificial-intelligence-index



2. AI Trends

1 AI Trends

Generative AI

Multi-modal AI

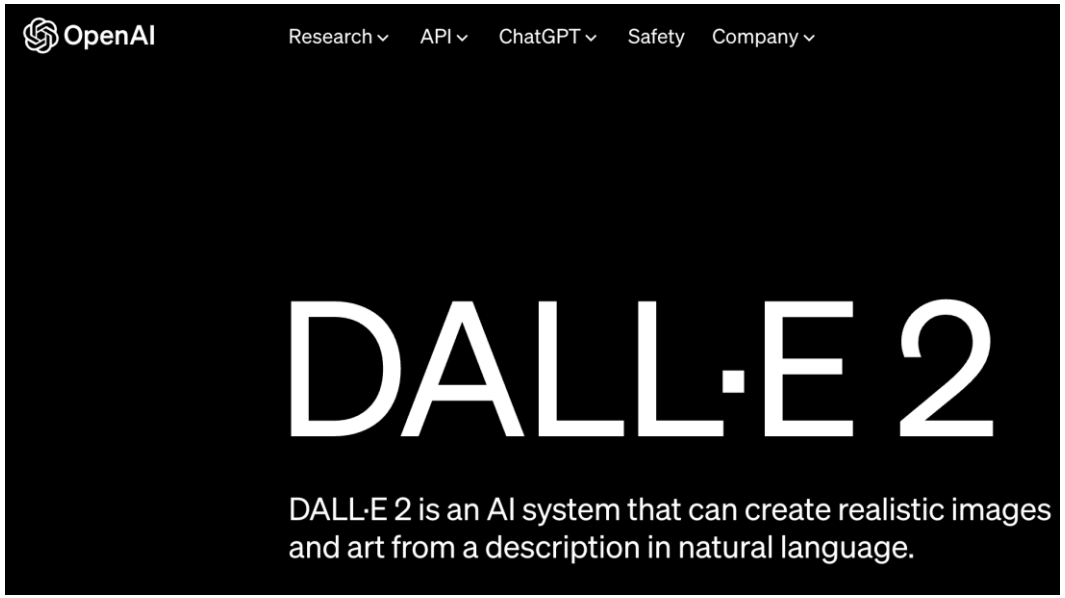
Trustworthy AI

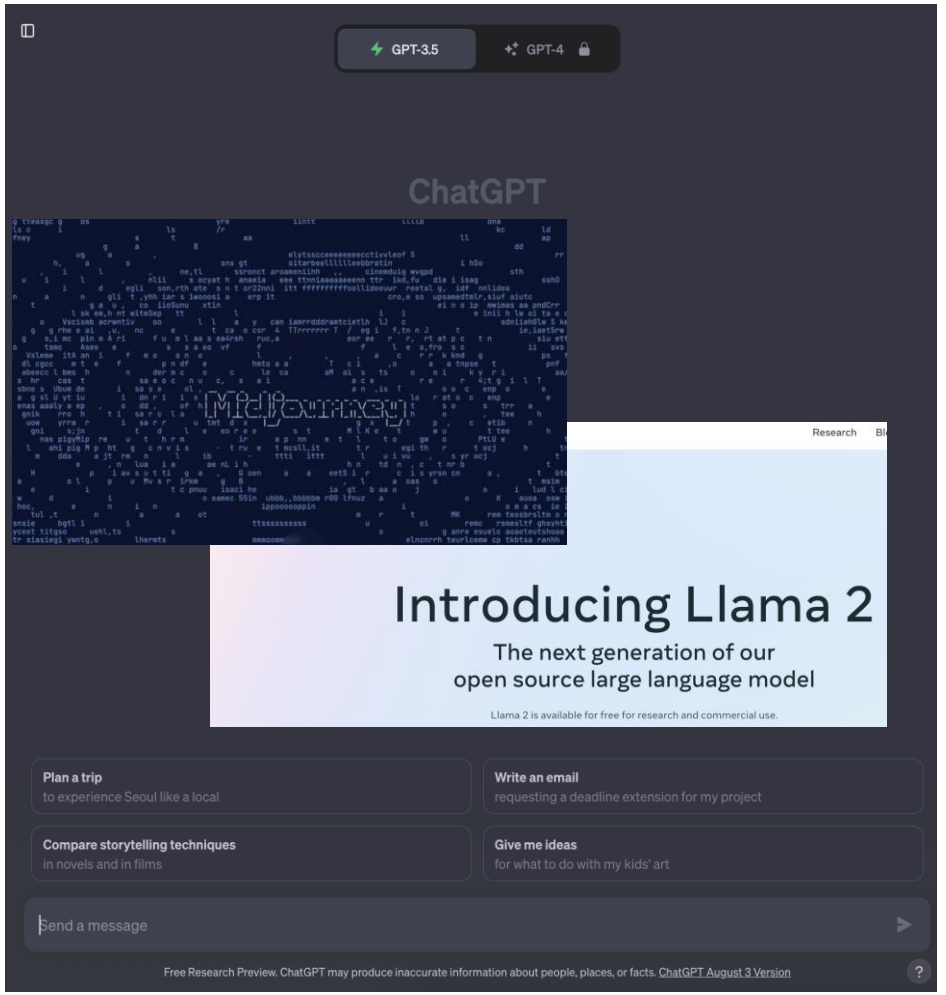
Data-centric AI

2

Generative AI

- **ChatGPT** popularized the Generative AI wave
- "**Prompt engineering**"





Hewlett Packard Enterprise | NVIDIA | DX 동국시스템즈

HPE - NVIDIA - 동국시스템즈

AI Solution Day

3 Going Multi-modal

- **Autonomous driving**: cameras, radars, lasers, lidar, etc.
- **Robotics**
- Multi-modal Text-to-Image
- Image-to-Text: Many captioning
- Any-to-Any: CoDi (Microsoft), Gato (Google)

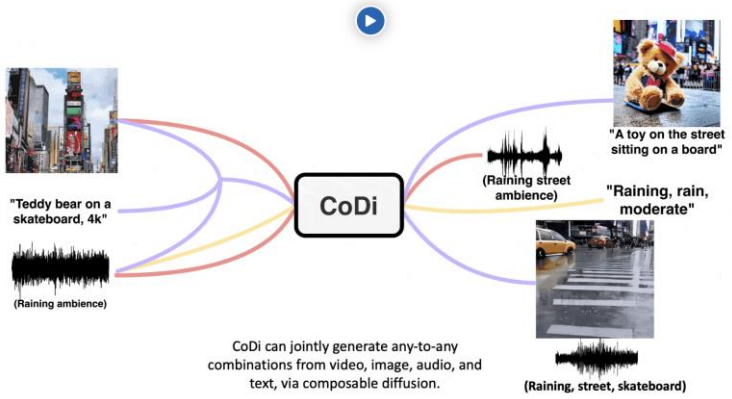


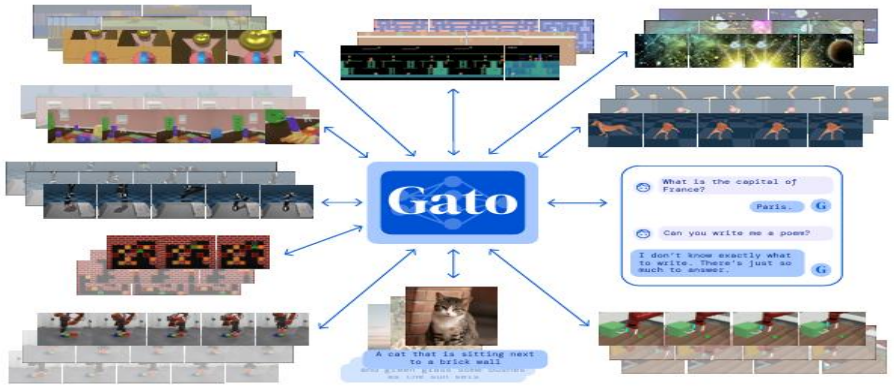
Image Source: https://www.microsoft.com/en-us/research/blog/breaking-cross-modal-boundaries-in-multimodal-ai-introducing-codi-composable-diffusion-for-any-to-any-generation/

TESLA'S AUTOPILOT DEPENDS ON A DELUGE OF DATA

But can a fire-hose approach solve self-driving's biggest problems?

Most companies working on automated driving rely on a small fleet of highly instrumented test vehicles, festooned with high-resolution cameras, radars, and laser-ranging lidar devices. Some of these have been estimated to generate 750 megabytes of sensor data every second, providing a rich seam of training data for neural networks and other machine-learning systems to improve their driving skills.

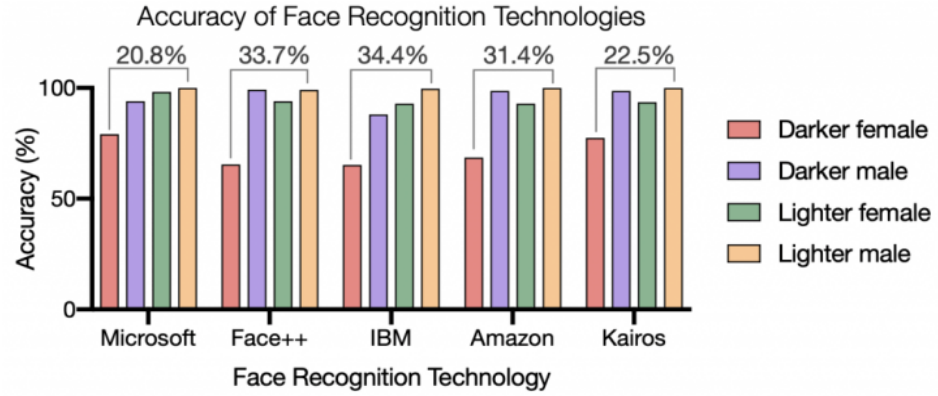
https://spectrum.ieee.org/tesla-autopilot-data-deluge

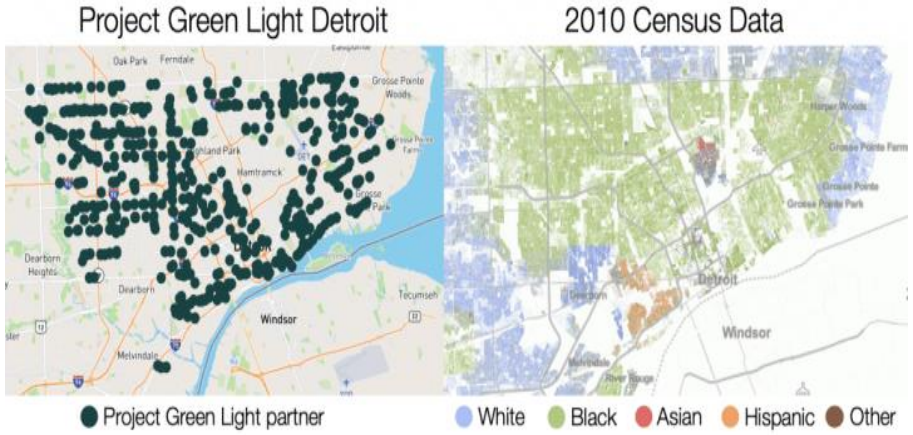


https://arxiv.org/pdf/2205.06175.pdf

4 Trustworthy AI

- **AI Bias (편향성)**
  - 소수 인종의 낮은 안면 인식 정확도
- **AI Data**에 입각한 우범지역 경찰순찰
- **Deepfake**:
  - Harrison Ford in Indiana Jones 5
  - 40 years younger?





https://sitn.hms.harvard.edu/flash/2020/racial-discrimination-in-face-recognition-technology/

Data-Centric: One model to rule them all?

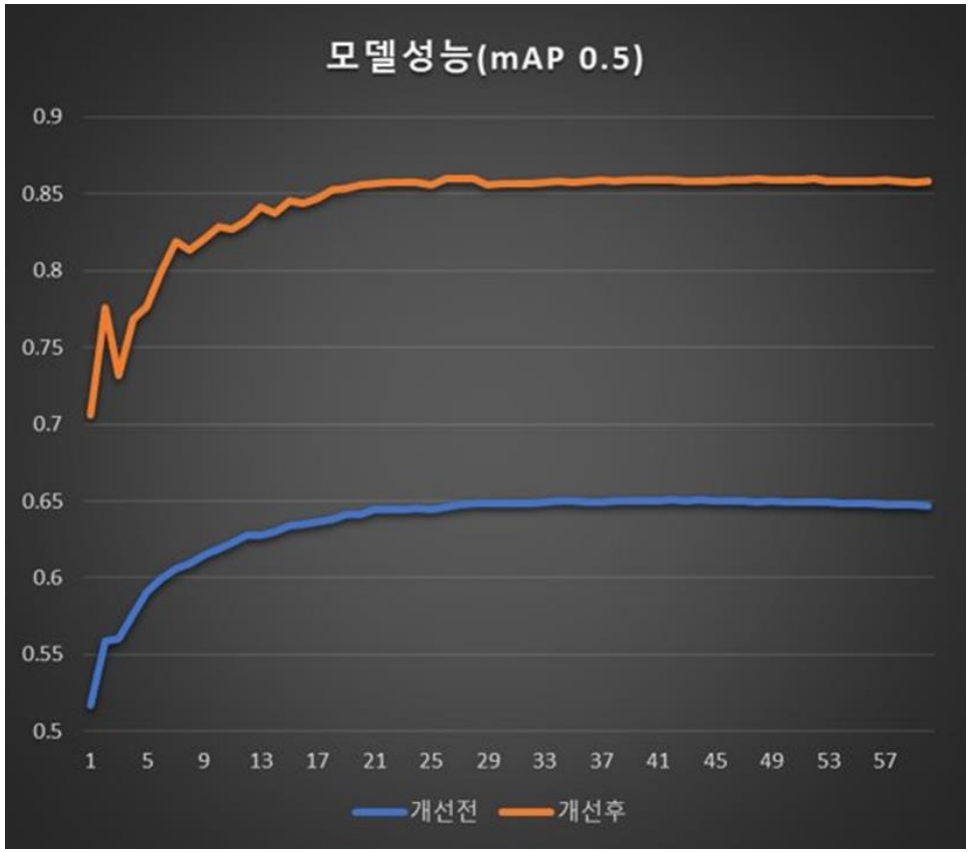
- Model 은 Transformer로 통합 되는 추세 (Foundation Models)
- Model architecture 보다 parameter 개수와 데이터의 종류와 양

AI Model	Parameters	Data
GPT1 (OpenAI)	117 million	7000 books
GPT2 (OpenAI)	1.5 billion	8 million web pages
GPT3 (OpenAI)	175 billion	45 TB text
DALL-E (GPT3) (OpenAI)	12 billion	Internet data (text + images)
PaLM (Google)	540 billion	780 B tokens of high-quality text
Wu Dao (悟道) (Beijing Academy of AI)	100 trillion	4.9 TB high quality images & text (English & Chinese)
ChatGPT3.5 (OpenAI)	175 billion	570 GB (300 billion words)
Llama 2 (Meta)	70 billion	2 trillion tokens

6

Data-Centric: A Tale of Two Datasets

- Same model (YOLOv5)
- Same raw data (AI Hub dataset): 30,000 images
- Different labels: AI Hub labels vs. Testworks fixed labels
- Training results show 43**%** (49% vs. 92% mAP) differences



3. Strategies for Quality Data

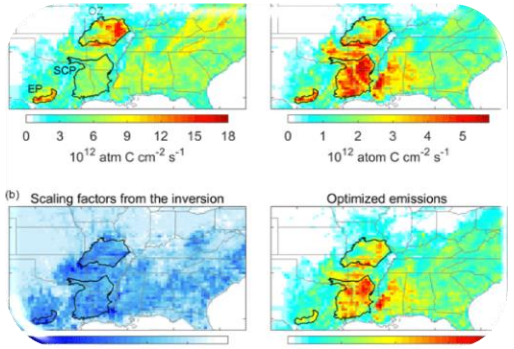
1

Strategies for better data

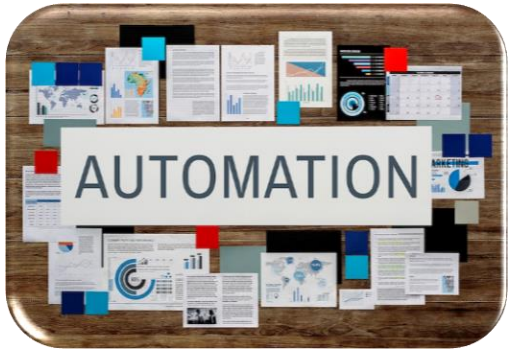
1. Divide and conquer [단계적 접근]
2. Quantify [정량화]
3. Visualize [시각화]
4. Automate [자동화]

Divide & Conquer

Quantify



Visualize

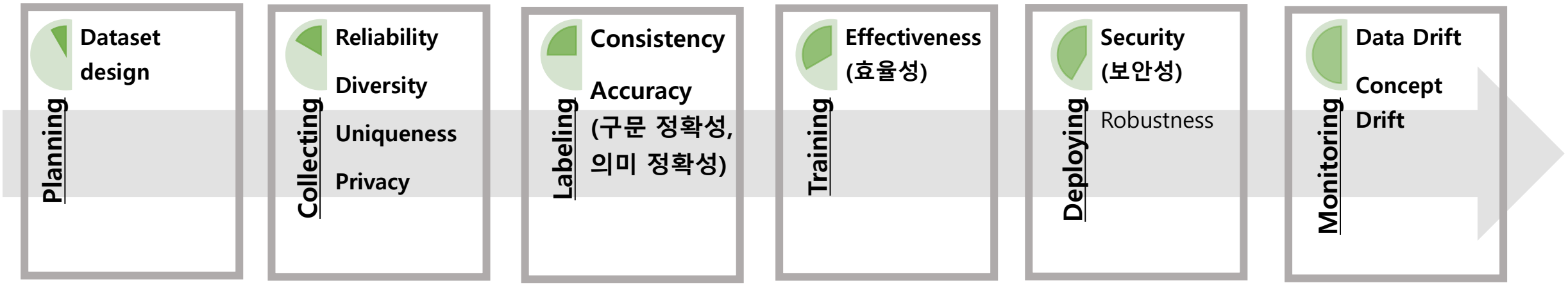


Automate

2

Strategy 1: Divide and conquer

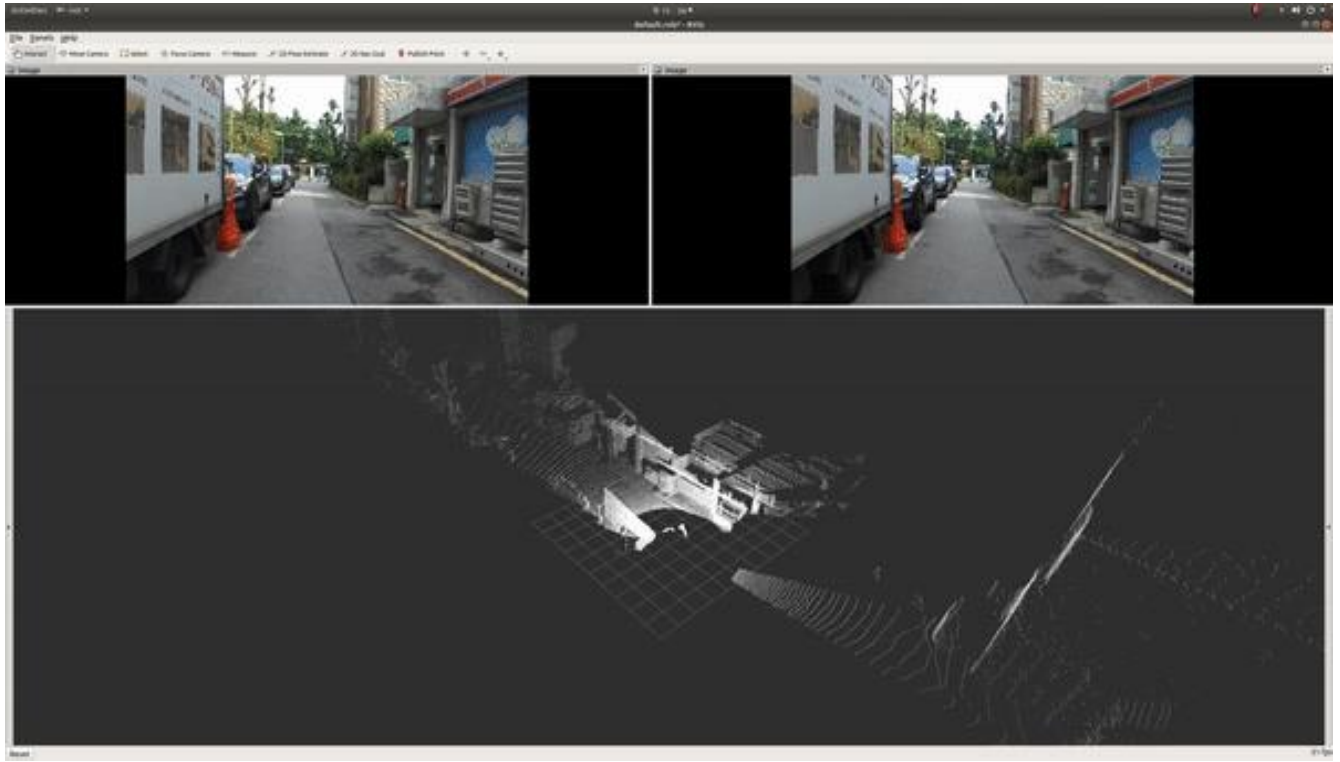
1. **Better Quality** → Build quality in ML Pipeline stages (ML Pipeline상에서 단계적 접근)
2. **Data quality** → Data quality criteria
3. **품질은 수집 단계에서부터**



3

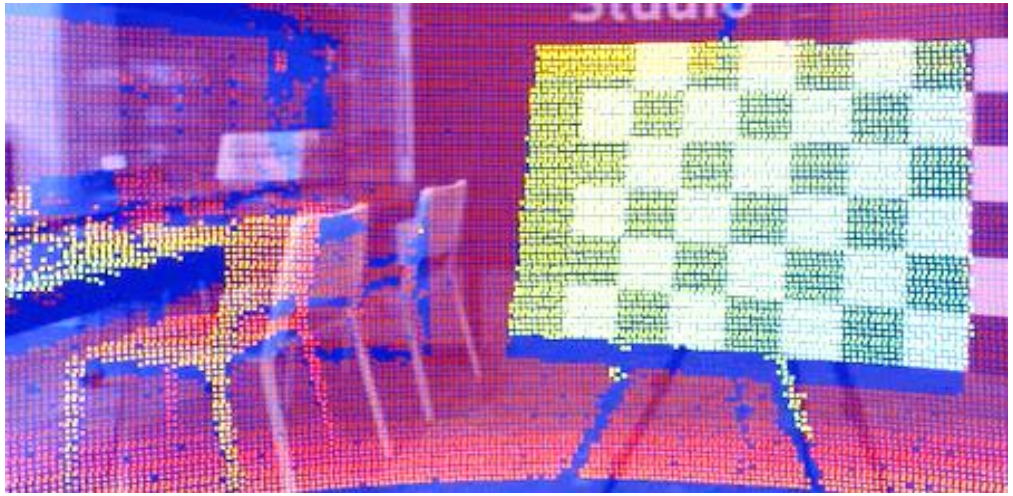
Collection strategy: synchronization & calibration

- Multi-modal data 구축 관건: 품질은 수집 단계에서부터
  - Synchronization (동기화), Calibration

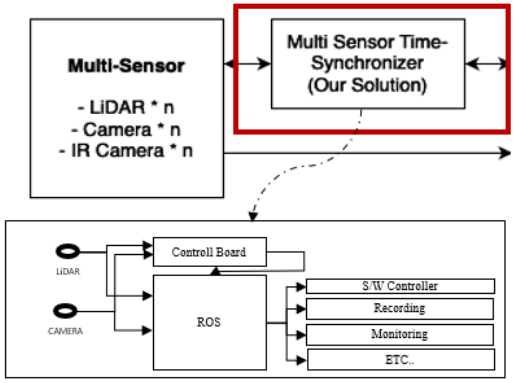


4 Collection strategy: synchronization

❑ Multi-modal data 구축 관건: **Synchronization**



시간 동기화가 되지 않은 경우

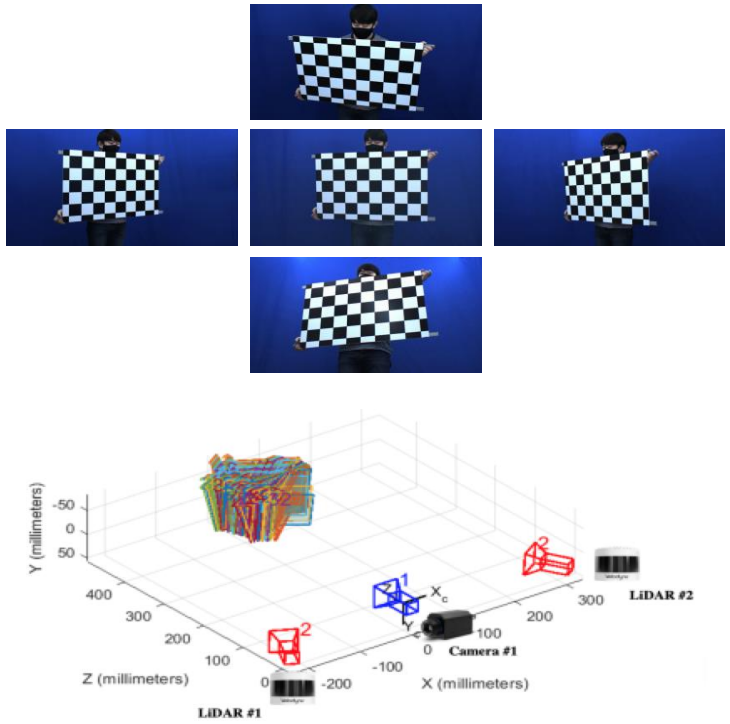


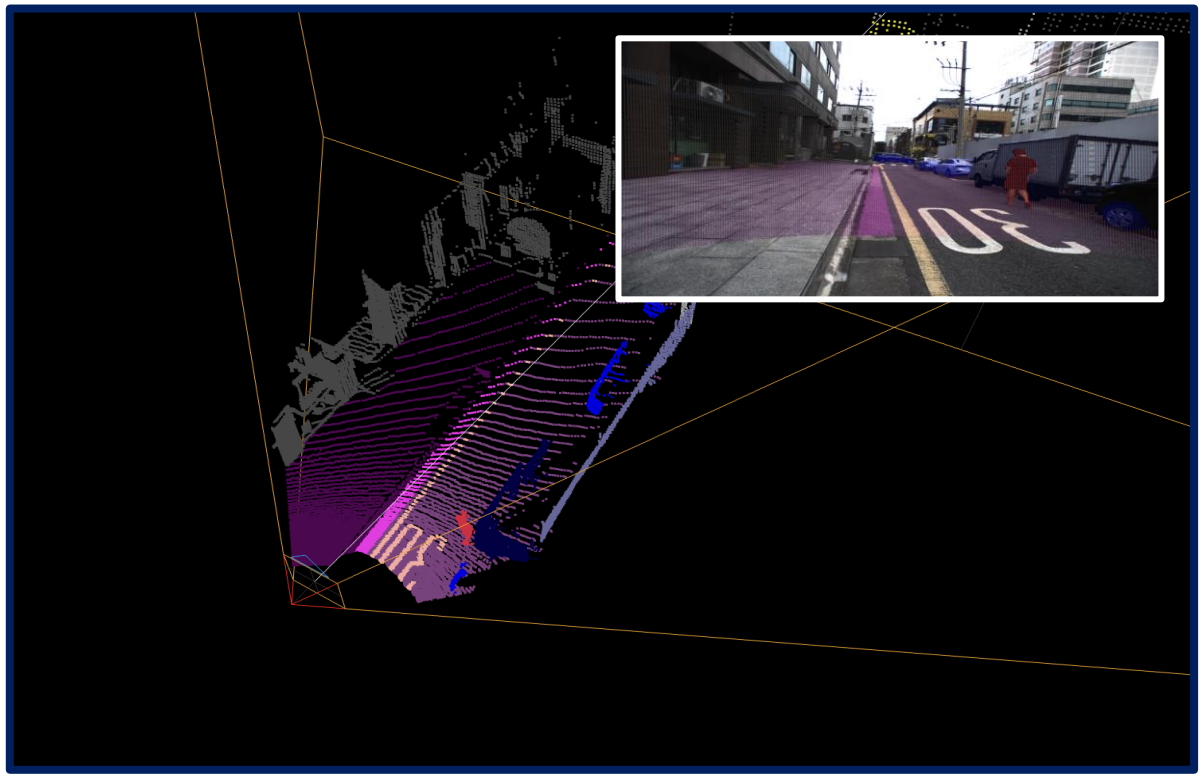
시간 동기화 이후

5

Collection strategy: calibration

❑ Multi-modal data: Calibration





6

Collection strategy: Calibration

❑ **Multi-modal data: Calibration 특허 기술**
(3차원 라이다 간 정합기술)


특허증
CERTIFICATE OF PATENT

특 허 Patent Number 제 10-2525568 호

출원번호 Application Number 제 10-2022-0164917 호

출원일 Filing Date 2022년 11월 30일

등록일 Registration Date 2023년 04월 20일

발명의 명칭 Title of the Invention
반사판을 이용한 원거리 라이다 위치 보정 시스템 및 방법

특허권자 Patentee
주식회사 테스트웍스(110111-******)
서울특별시 송파구 백제고분로41길 40 , 401호(송파동, 부영빌딩)

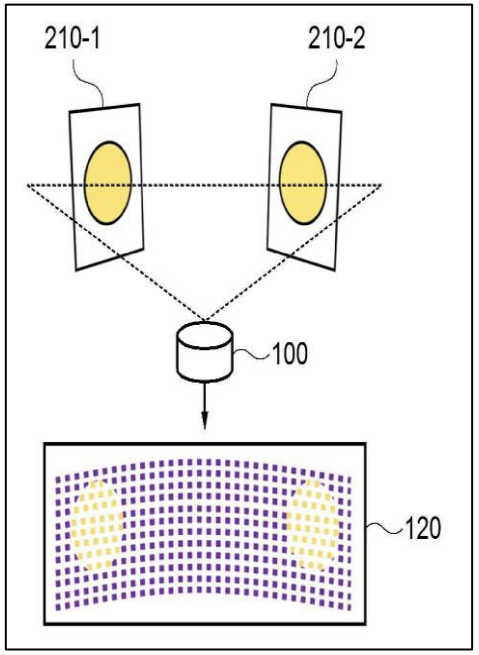
발명자 Inventor
등록사항란에 기재

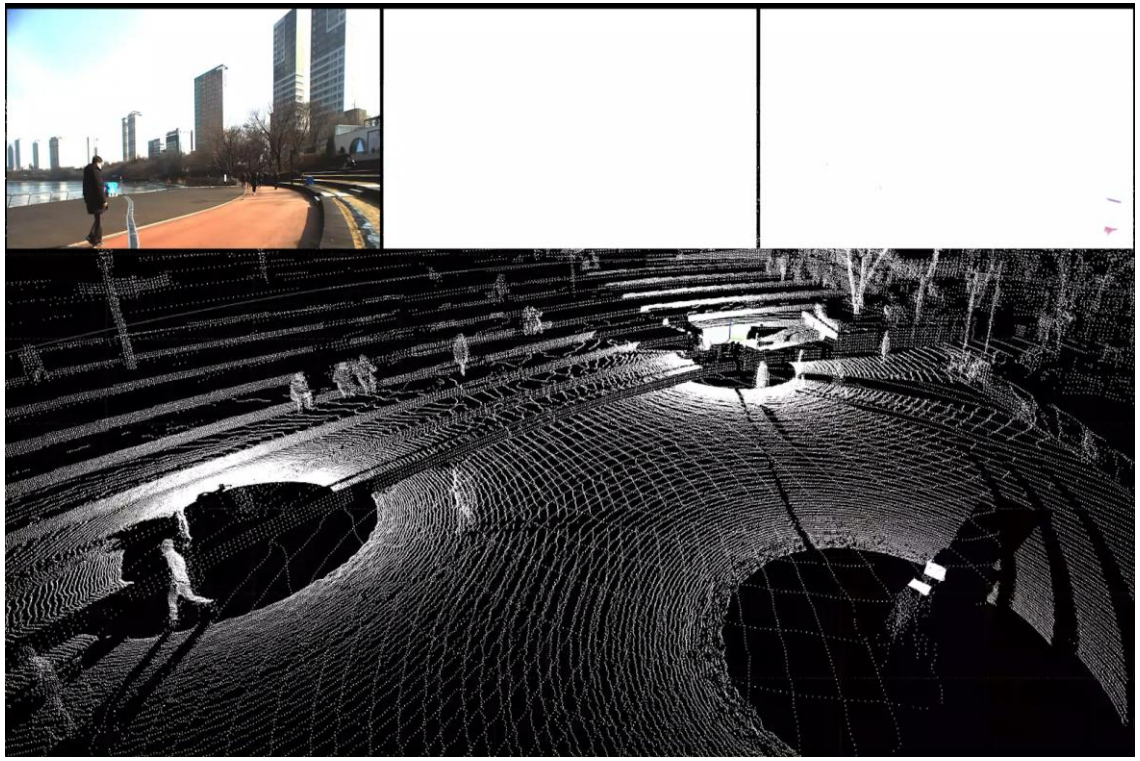
위의 발명은 「특허법」에 따라 특허원부에 등록되었음을 증명합니다.
This is to certify that, in accordance with the Patent Act, a patent for the invention has been registered at the Korean Intellectual Property Office.

2023년 04월 20일

특허청장
COMMISSIONER,
KOREAN INTELLECTUAL PROPERTY OFFICE

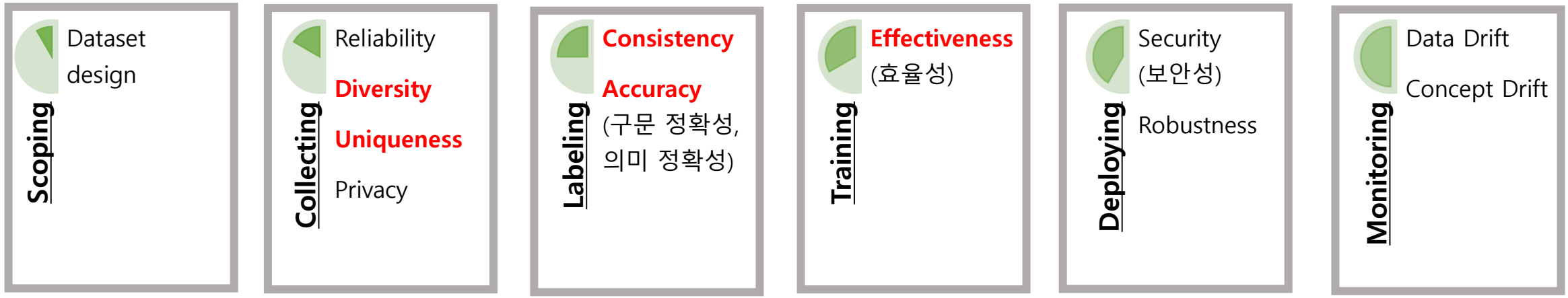
특허청
Korean Intellectual Property Office





Strategy 2: Quantify

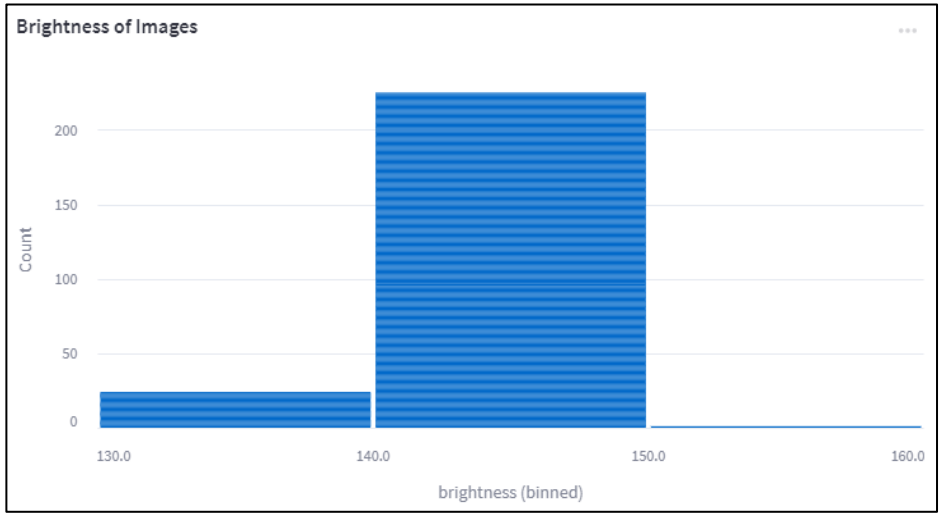
1. Turn quality to quantity (품질의 정량화)
2. Use metrics to measure quality

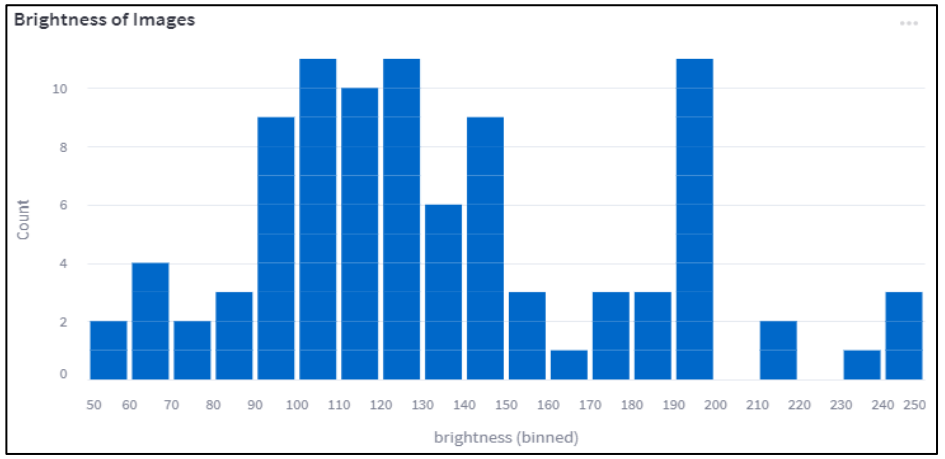


Strategy 2: Quantify

- Diversity criteria need to be quantified:
  - Brightness – Balanced or imbalanced?
  - Aspect ratio?
  - Class distributions?
  - File sizes?
  - Etc.

사진의 음영





Strategy 2: Quantify

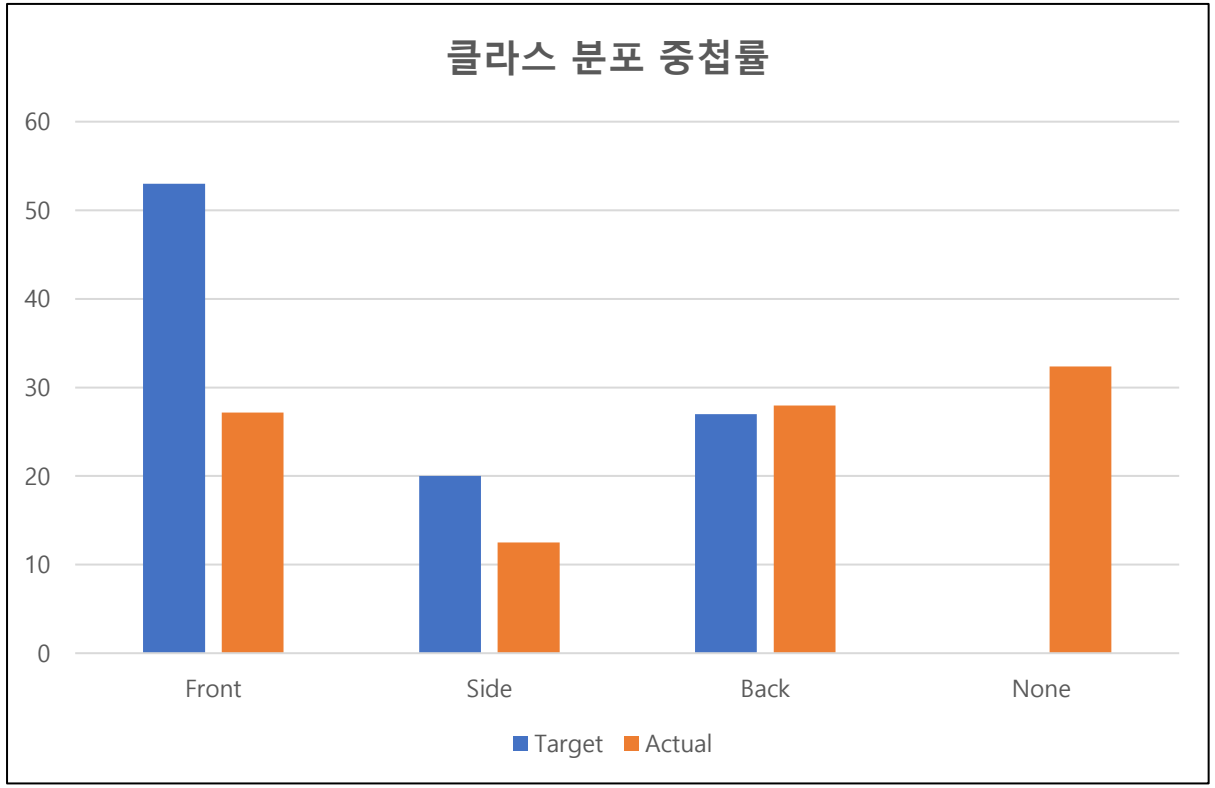
예) Diversity (다양성) => 구성비 중첩률

· 다양성 항목 : IoU 50% 이상
· 목표대비 결과 비율

구분	목표	결과	중첩률(IoU)
질환	40%	50%	81.8%
정상	60%	50%	

· 검사 결과 구성비 중첩률(%) = (40 + 50) / (50 + 60) * 100 = 81.8

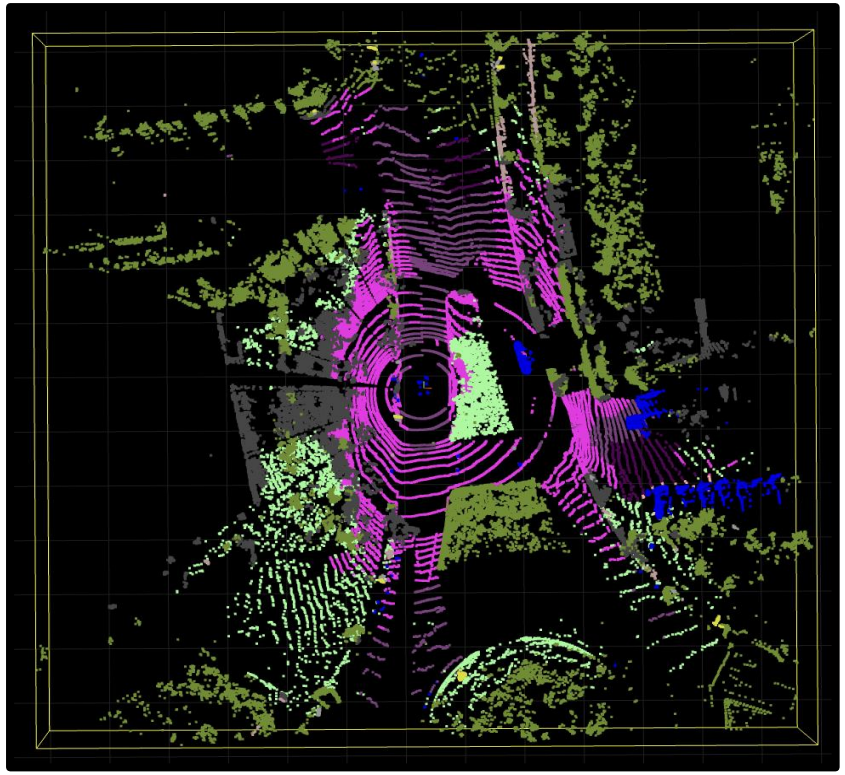
인공지능 학습용 데이터 품질관리 가이드라인 v3.0 p. 151

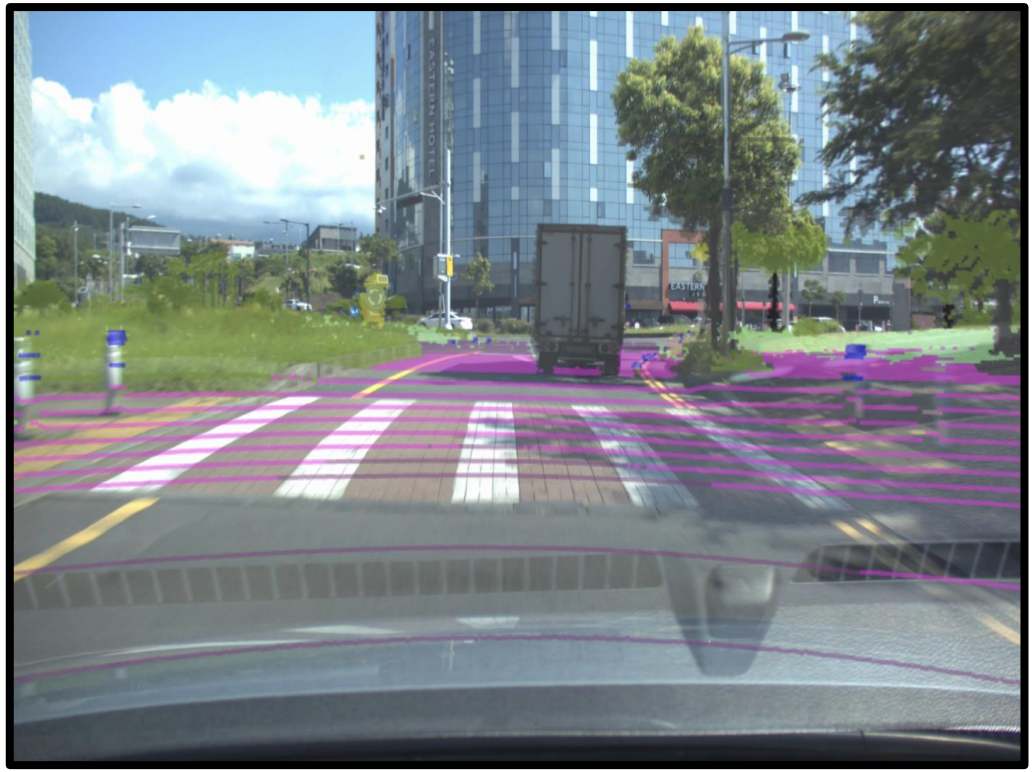


중첩률 = 50%

Strategy 2: Diversity

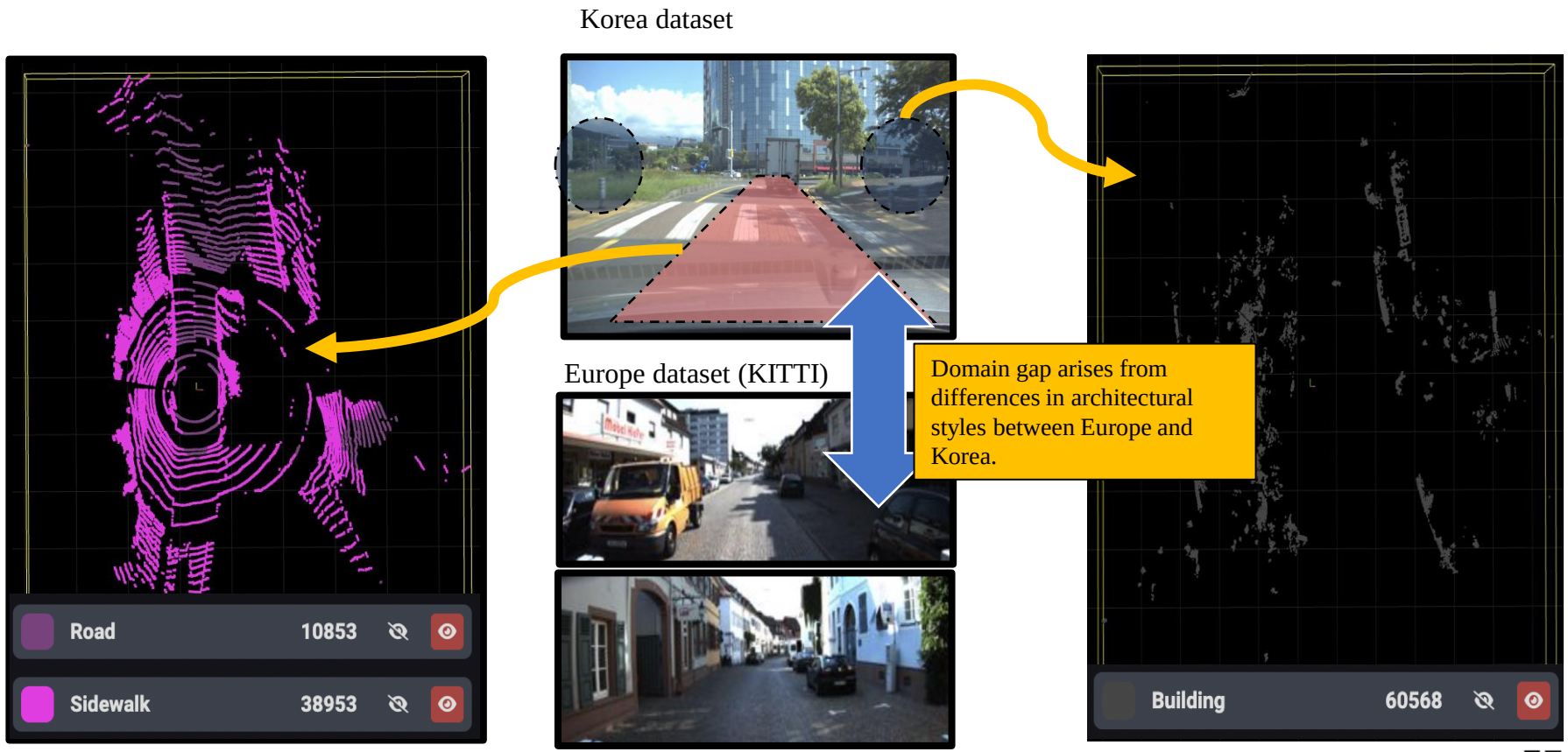
- Semantic segmentation auto-labeling using [The KITTI Benchmark Dataset]





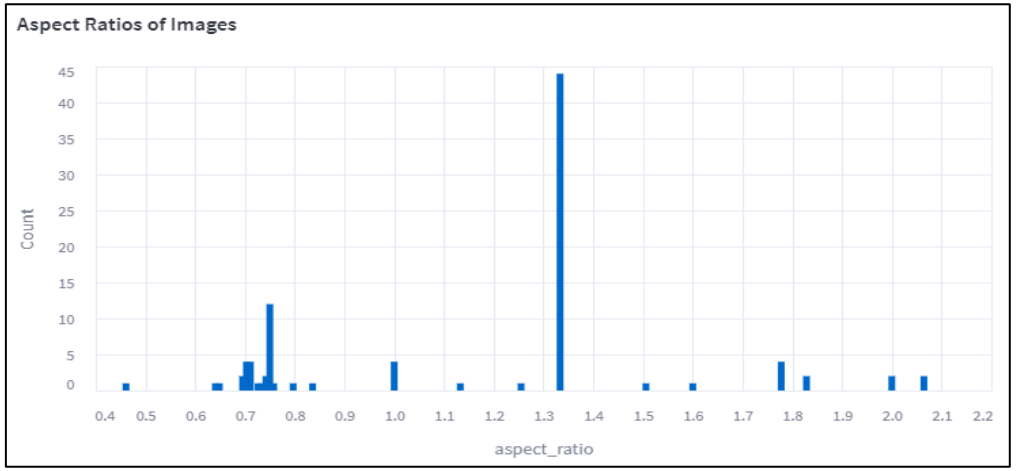
Strategy 2: Diversity

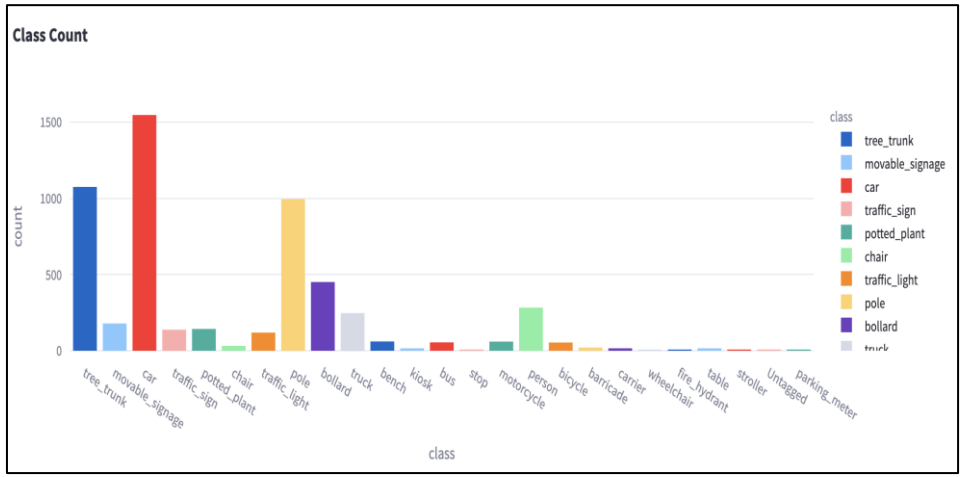
- ❑ Auto-labeling error due to different architectural styles (domain gap)
- ❑ The importance of data quality is re-enforced

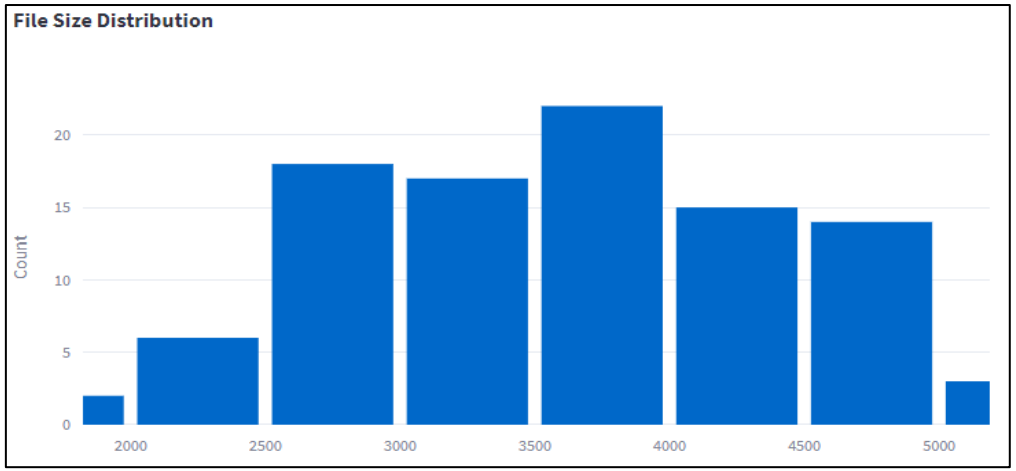


Strategy 3: Visualize

- Visualize different aspects of a dataset
  - Class count
  - File size
  - Aspect ratio, etc.

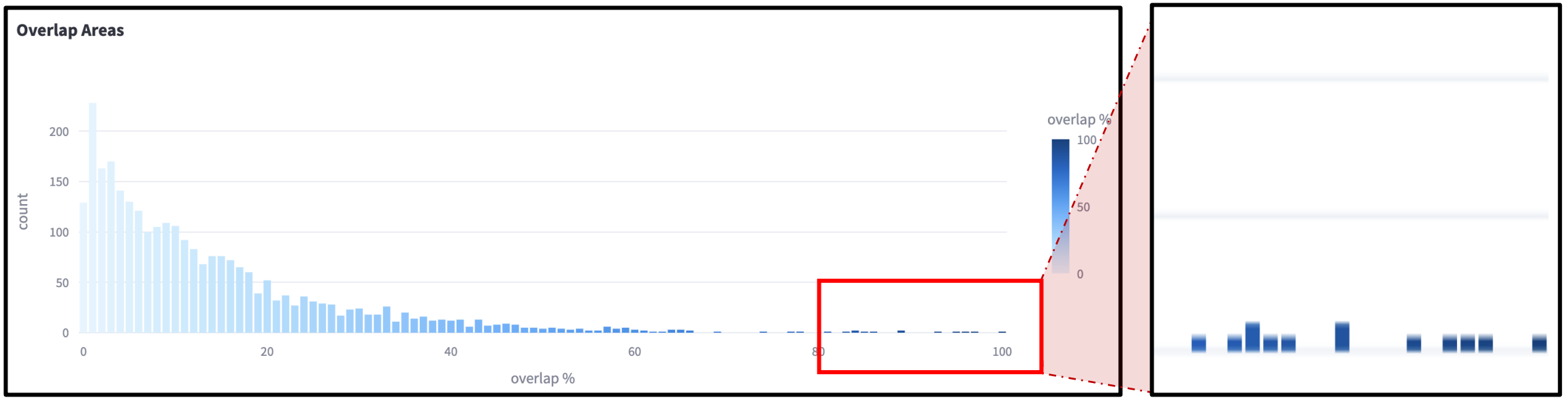






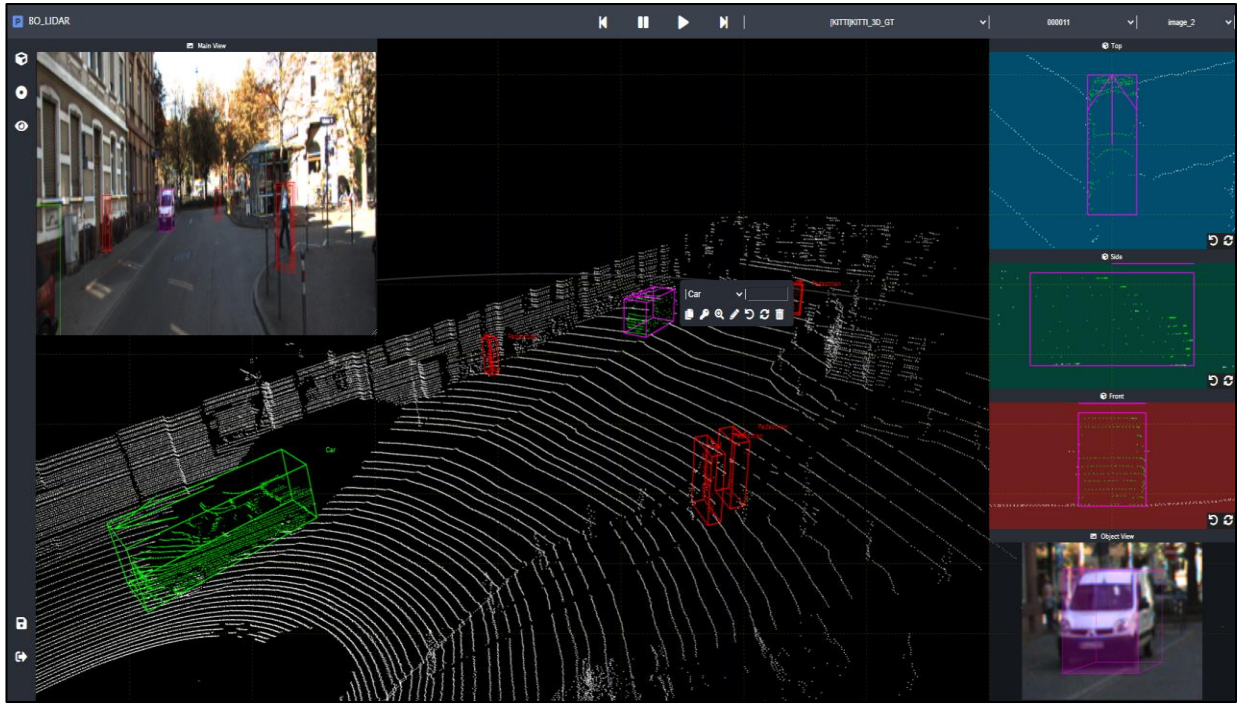
Strategy 3: Visualize

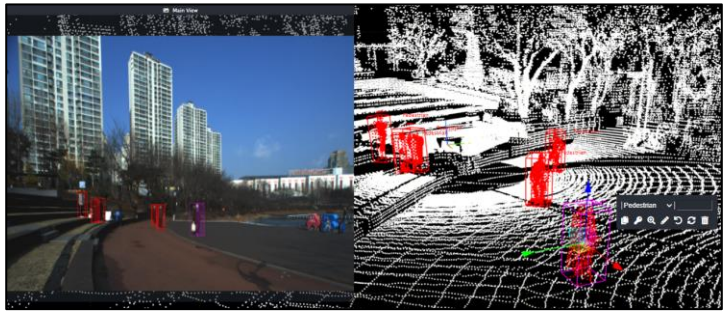
- Overlapping labels (중첩된 라벨)
  - Tool bugs?
  - Human errors?



Strategy 3: Visualize

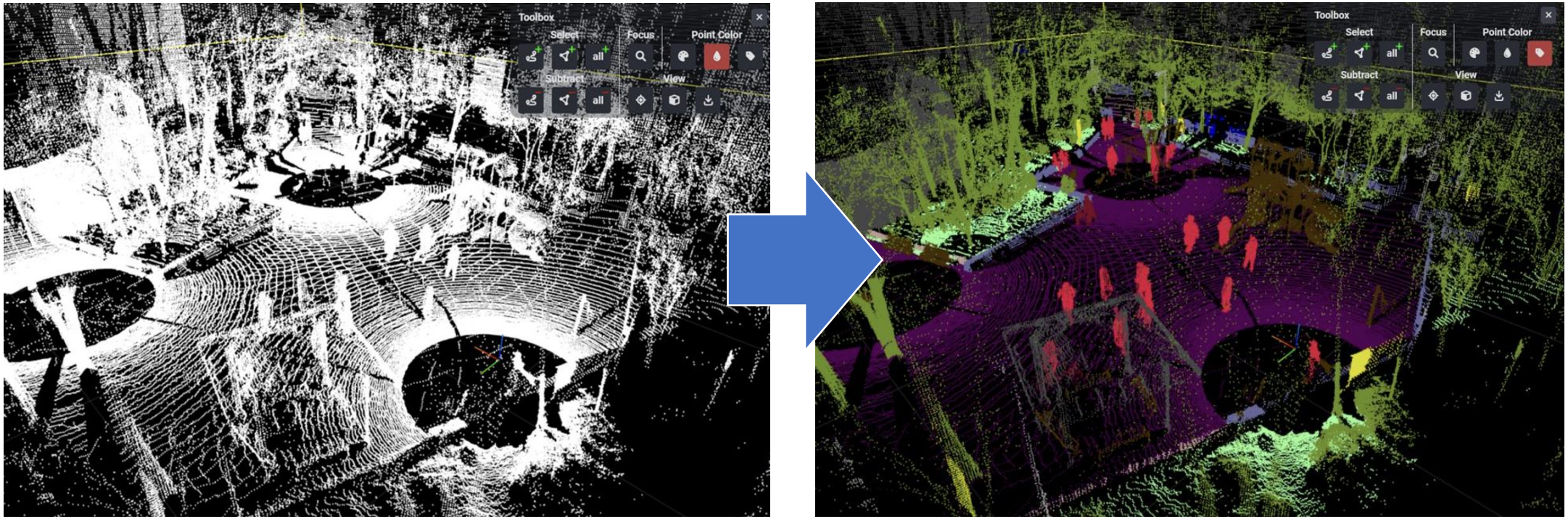
- ❑ Multi-modal data labeling – Multi-modal Annotation Tool (blackolive 3D)
- ❑ 2D RGB + 3D Point Cloud data





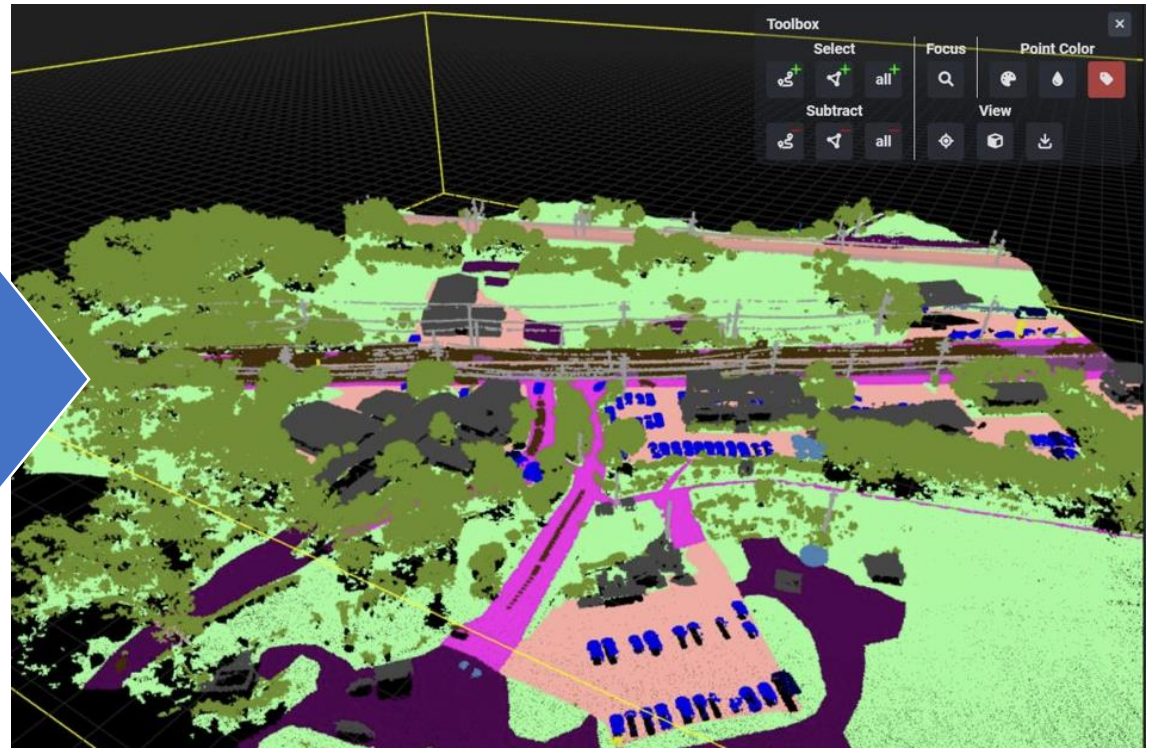
Strategy 3: Visualize 3D Segmentation

- ❑ blackolive 3D, Testworks Multi-LiDAR Dataset (LiDAR + Segmentation) 가공
- ❑ Raw data: LiDAR + Mono



Strategy 3: Visualize 3D Segmentation

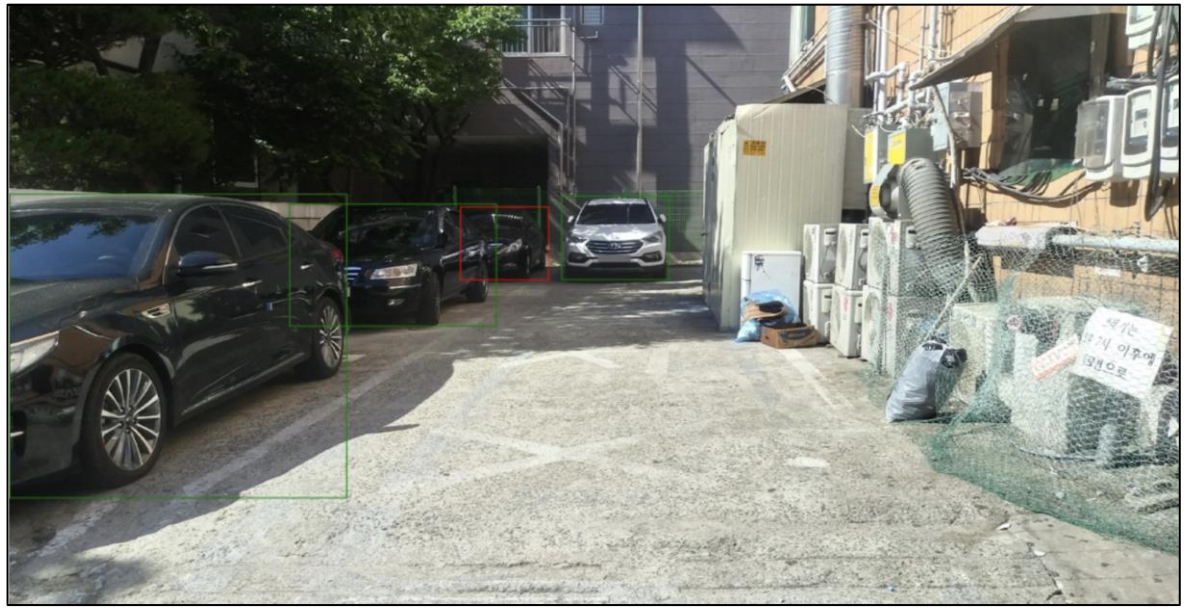
- blackolive 3D, Testworks Multi-LiDAR Dataset (LiDAR + Segmentation) 가공
- Raw data: LiDAR + Mono

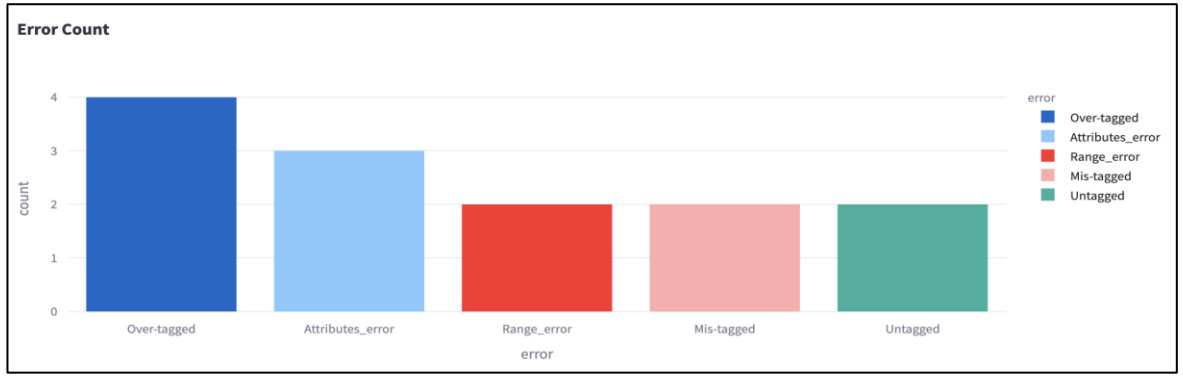


Strategy 4: Automate

❑ **Automate + Manual Review**:

- Auto Labeling
- Auto Reviewing:
  - Duplicate labels
  - Overlapping labels
- 자동 라벨/검수 후 확인 수정



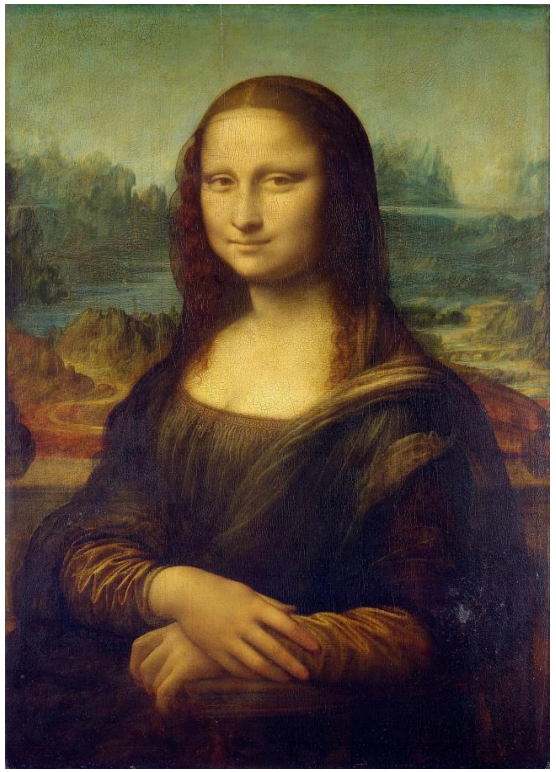


4. Conclusions

1

Trustworthy revisited

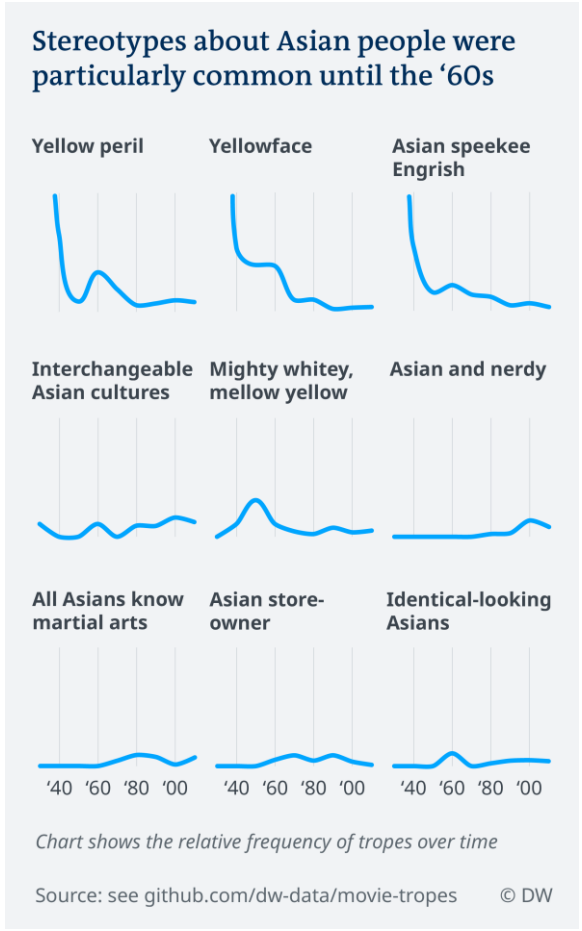
- Mis-use of technology
- Use of camera & computers by Nazis for Holocaust

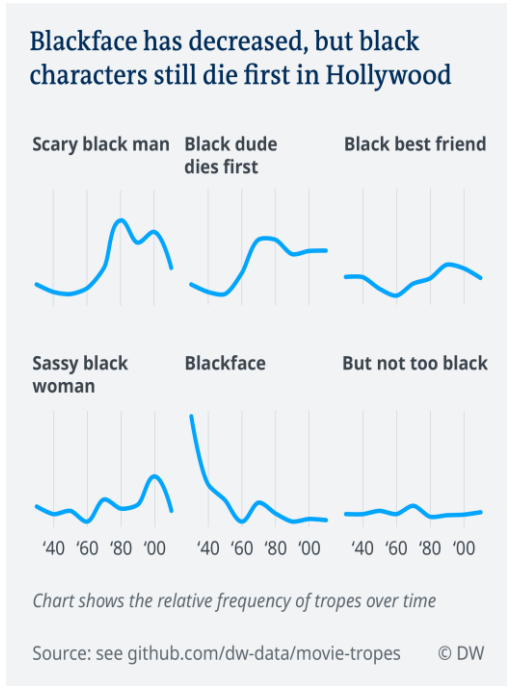


Camera

2 Stereotypes

- Asian stereotypes in America?
  - Short?
  - Good at math?
  - Bad drivers?
  - Playing the piano & the violin?
- What are your stereotypes?
  - Americans
  - Koreans
- How difficult is it to change stereotypes?





Source: https://www.dw.com/en/hollywood-movies-stereotypes-prejudice-data-analysis/a-47561660

3 The Ugly Model

- Stereotypes are hard to change
- White privilege => sets the beauty standard
- Asians = ugly

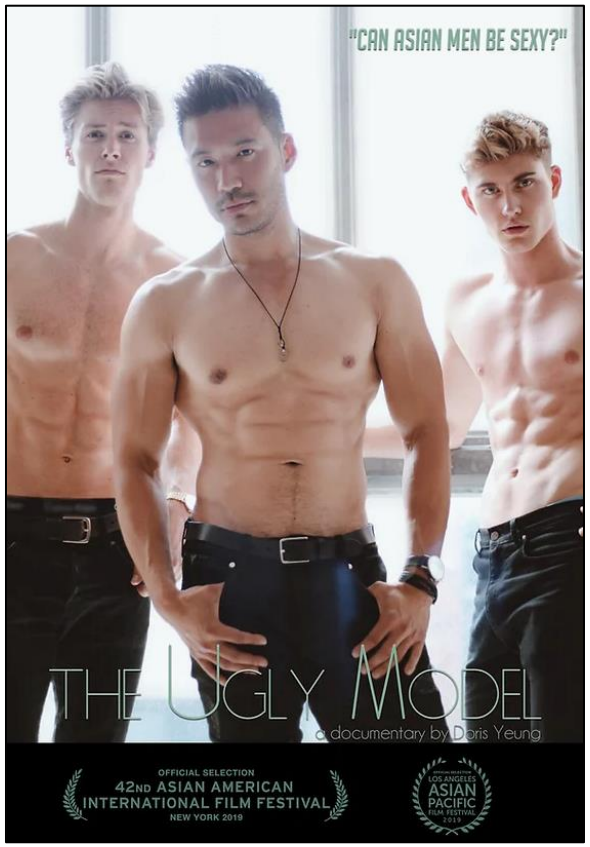


Image Source: https://dearstraightpeople.com/2018/11/13/ugly-model-kevin-kreider/

4 Strategies for Quality AI Data

❑ **AI can be quantified, visualized and improved with right strategies**

- Easier to change than human biases and stereotypes
- 사람보다 개선하기 쉬운 AI

❑ **Strategies for quality AI Data**

- Divide and conquer: 품질은 수집 단계에서부터
- Quantify: 품질 개선의 시작은 정량화
- Visualize: 품질을 보게 하라
- Automate: 품질의 자동 개선

Hewlett Packard Enterprise | NVIDIA | DX 동국시스템즈

HPE - NVIDIA - 동국시스템즈

AI Solution Day

Hewlett Packard Enterprise | NVIDIA | DX 동국시스템즈

HPE - NVIDIA - 동국시스템즈

AI Solution Day

THANK YOU