Theme

Challenges and Advantages of the Coexistence of Raw Data, Metadata, and Synthetic Data

InfoSys 2026 & InfoWare 2026

PANEL #2

VALENCIA
March 2026

Moderator

Prof. Dr. Andreas Schmidt, Karlsruhe Institute of Technology &
University of Applied Sciences, Karlsruhe,
Germany

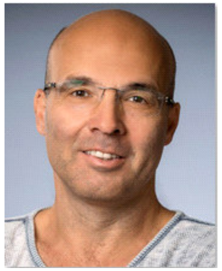
Panelists

Emeritus Prof. Dr. Malcolm Crowe, University of the West of Scotland,
UK, Scotland

Prof. Dr. Steve Zhou, University of Houston Downtown,
USA

Prof. Dr. Constantine Kotropoulos, Aristotle University of Thessaloniki,
Greece

Definitions



Andreas Schmidt

Karlsruhe Institute of Technology

&

University of Applied Sciences Karlsruhe

Challenges and Advantages of the Coexistence of Raw Data, Metadata, and Synthetic Data

- **Raw data**: Data in it’s original form - not processed, formatted, coded, or analyzed (sensor data, output from an experiment, ...)

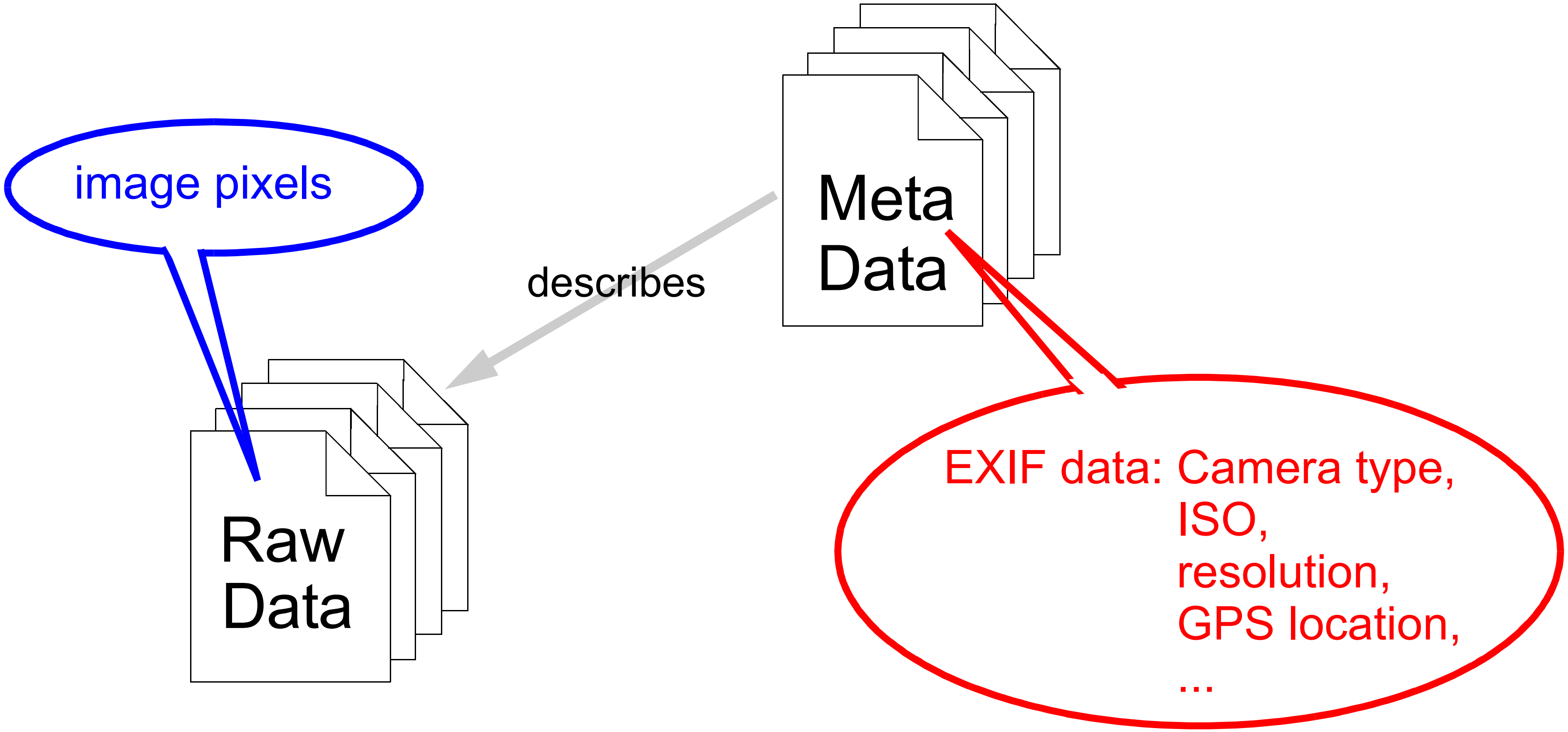
Content

- **Synthetic data**: artificially generated data not produced by real-world events. Typically created using algorithms it mimics the features, structures, and statistical properties of real world data

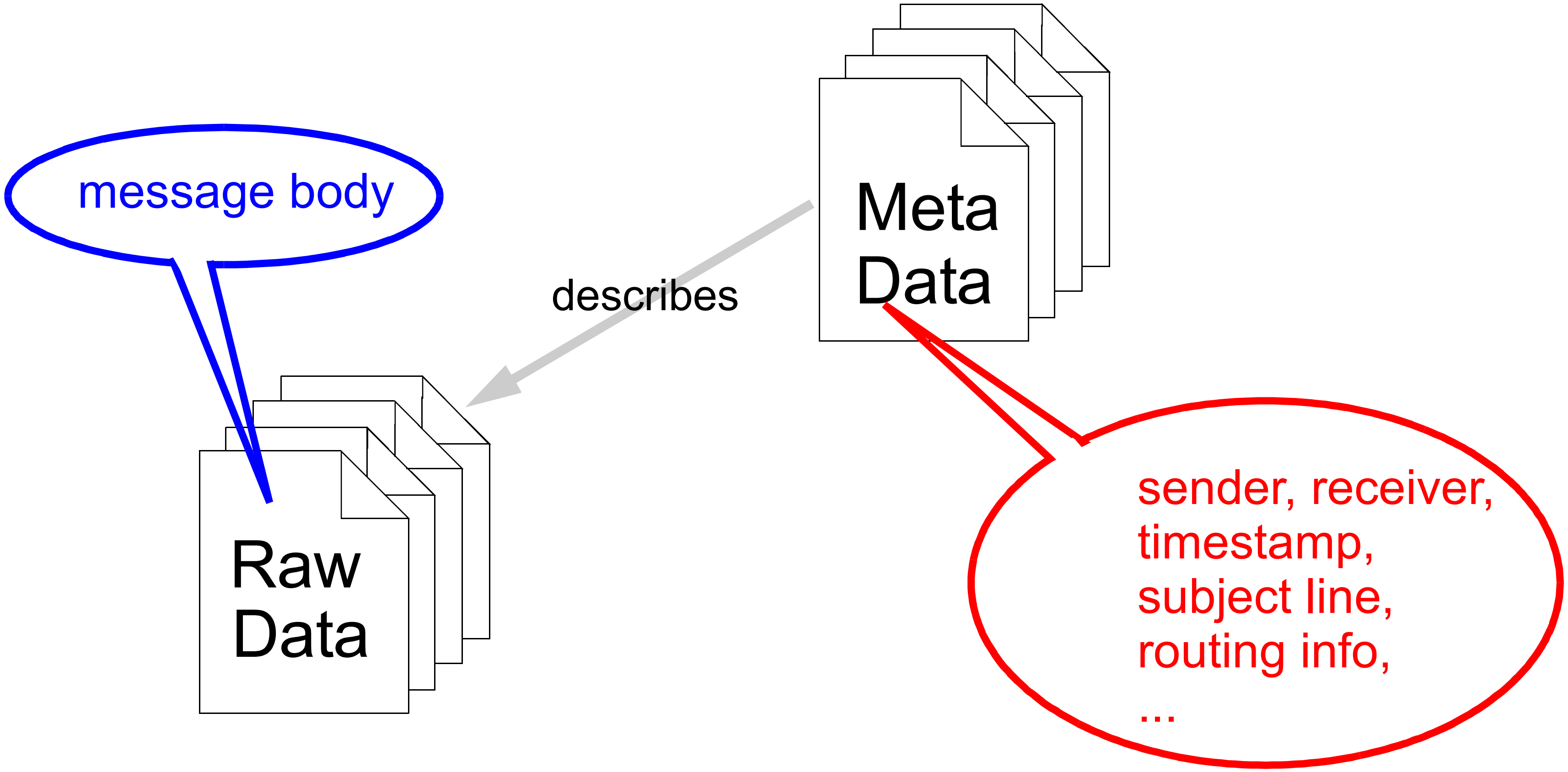
- **Meta data**: „data about data“ (data that defines and describes the characteristics of other data)

Context

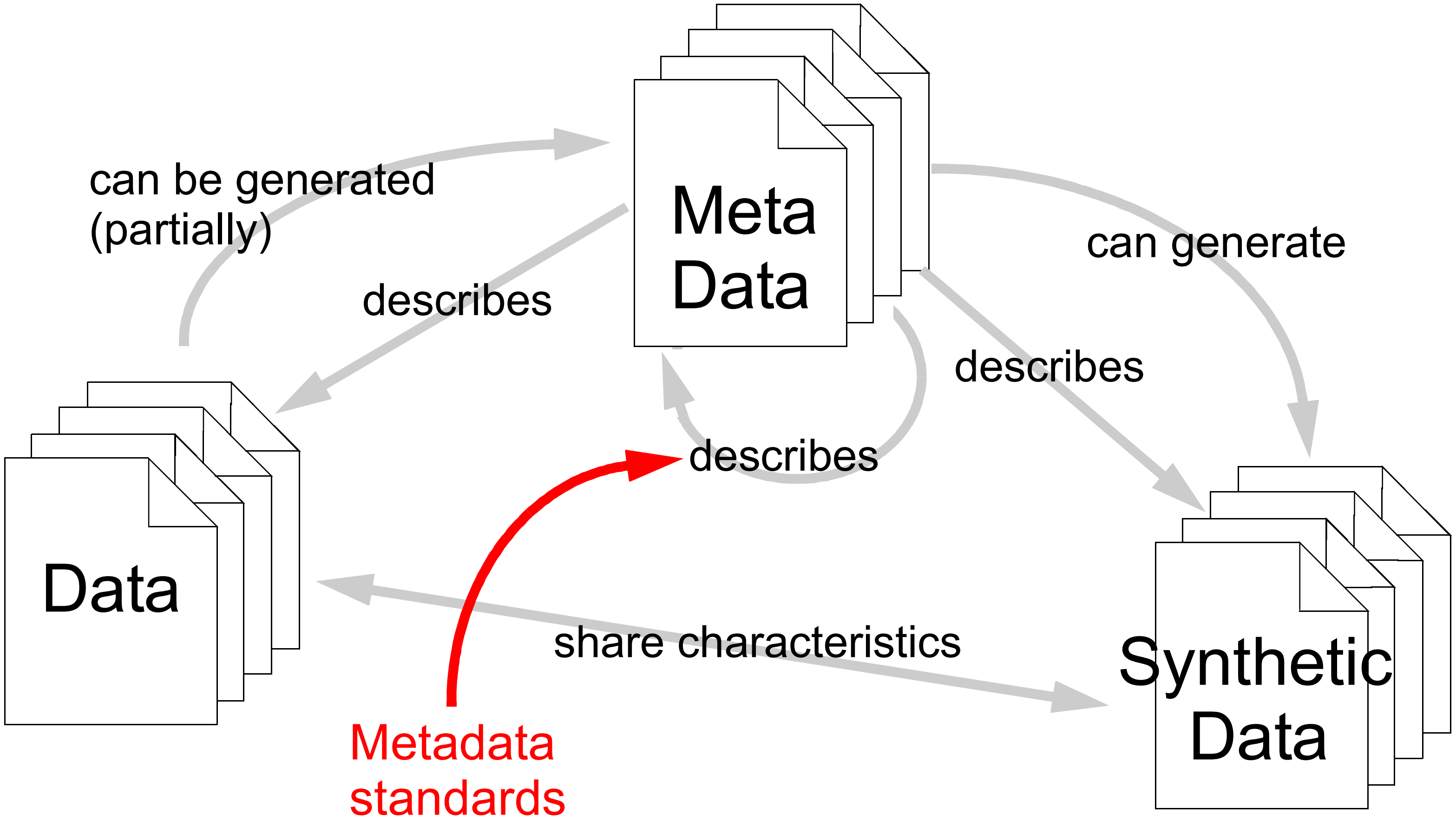
Example: Photography



Example: Email


message body
describes
Meta Data
Raw Data
sender, receiver, timestamp, subject line, routing info, ...

Relationship between ...



Panelist Position

... and now

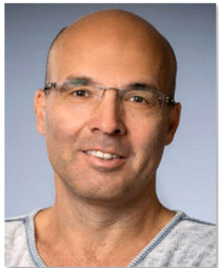
the panelists positions ....

IARIA

Panelist Position

VALENCIA
March 2026

Metadata in RDBMS



Andreas
Schmidt

Karlsruhe
Institute of
Technology

&

University
of Applied
Sciences
Karlsruhe

Codd's 12 Rules in DBMS [1] ...

- Rule 1: The Information Rule

  All information in a relational data base is represented explicitly at the logical level and in exactly one way — by values in tables.

- Rule 4: Dynamic online catalog based on the relational model

  The database description is represented at the logical level in the same way as ordinary data, so that authorised users can apply the same relational language to its interrogation as they apply to the regular data.

[1] Codd's 12 rules. https://en.wikipedia.org/wiki/Codd%27s_12_rules

Metadata in RDBMS

- Set of read-only tables and views, that provide information about the database
- Provides information about ...
  - Databases, Users/Privileges, Tables, Views, Columns, Constraints, Statistics, Indexes, Source code, Physical implementation, ...
- Access via SQL
- Use cases:
  - IDE usage
  - Query optimization
  - Scripting (SQL to generate SQL)
  - Code generation
  - Extraction of Domain Model
  - Construction of data-browsers
  - Auto-completion
  - Data governance
    - monitoring of compliance with legal regulations
    - Improving data/schema quality
  - ...

Panelist Position

VALENCIA
March 2026

- Raw Data = measured observations, actual events
  - No enhanced or altered copies: provenance, curation
  - Big business very rarely allows this, they want spin, context, explanation
- Metadata = how, where and when the data was obtained
  - Includes ethical and ownership aspects
- Synthetic data = the sort of thing we expected to get
  - Unconscious bias: what we see depends on what we expect

    The "Mirror of Nature": What aspects, what sample?
  - Some historians *(Toynbee)* take a lot of trouble to declare their personal bias/background (as metadata)
  - What we collect risks depending on what we would like to get (conscious bias)
- Can we keep these separate?

Malcolm
Crowe
UWS Scotland

Panelist Position

VALENCIA
March 2026

Steve Zhou
University of Houston Dowtown

- Data in Business: Supply Chain & Enterprise Applications
- Raw data in business
- Metadata in business
- Synthetic data in business

IARIA

Panelist Position

VALENCIA
March 2026

Common Sources & Examples

Transaction Logs
Sales, payments, service interactions

Survey Responses
Customer feedback, market research

Social Media
Posts, comments, reactions, engagement

Surveillance
Video feeds, security footage

Enterprise Data Collection Systems

ERP Systems
Financial, HR, manufacturing data

SCM Software
Inventory, logistics, supplier data

Web Analytics
User behavior, traffic patterns

CRM Platforms
Customer interactions, sales pipeline

Mobile Apps
Usage data, location services

Panelist Position

Business Metadata

Context & Meaning

Translates technical reality into business meaning, bridging the gap between how data is stored and what it actually represents.

- Business glossary terms and definitions
- Data ownership and stewardship
- Business rules and calculations
- Certified definitions and KPI logic

Usage Metadata

Consumption & Value

Captures how data is actually being consumed, turning metadata from passive documentation into active intelligence about data value and risk.

- Query patterns and access frequency
- Dashboard and report dependencies
- User ratings and certifications
- Data popularity and trust scores

Generation Methods

- Simulation: Crystal Ball, Arena, Anylogic
- Logistics/Network Optimization: CPLEX
- Statistical Programming: R, Matlab, Python
- AI: SDV tabular

Enterprise Use Cases

AI/ML Model Training

Train models when production data is limited or sensitive

Software Testing

Realistic test data without privacy risks

Product Development

Build features before production data exists

Data Sharing

Share insights without exposing real data

Sales Demos

Showcase capabilities without privacy exposure

IARIA

Panelist Position: Raw and Need for Synthetic Data

VALENCIA
March 2026

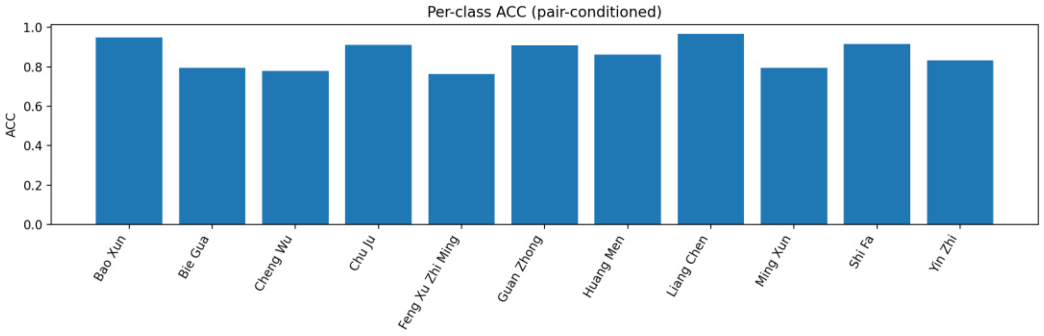
Panel #2 Challenges and Advantages of the Coexistence of Raw Data, Metadata, and Synthetic Data

- Minority classes /Data augmentation
  - Scribe verification
  - Speech Pathology Detection
- Generative adversarial networks and diffusion models for generating synthetic images
  - Forensics applications (Face aging/Detection of AI generated content)

Constantine
Kotropoulos
AUTH

Table 1. Data categories and number of samples in the training and test sets.

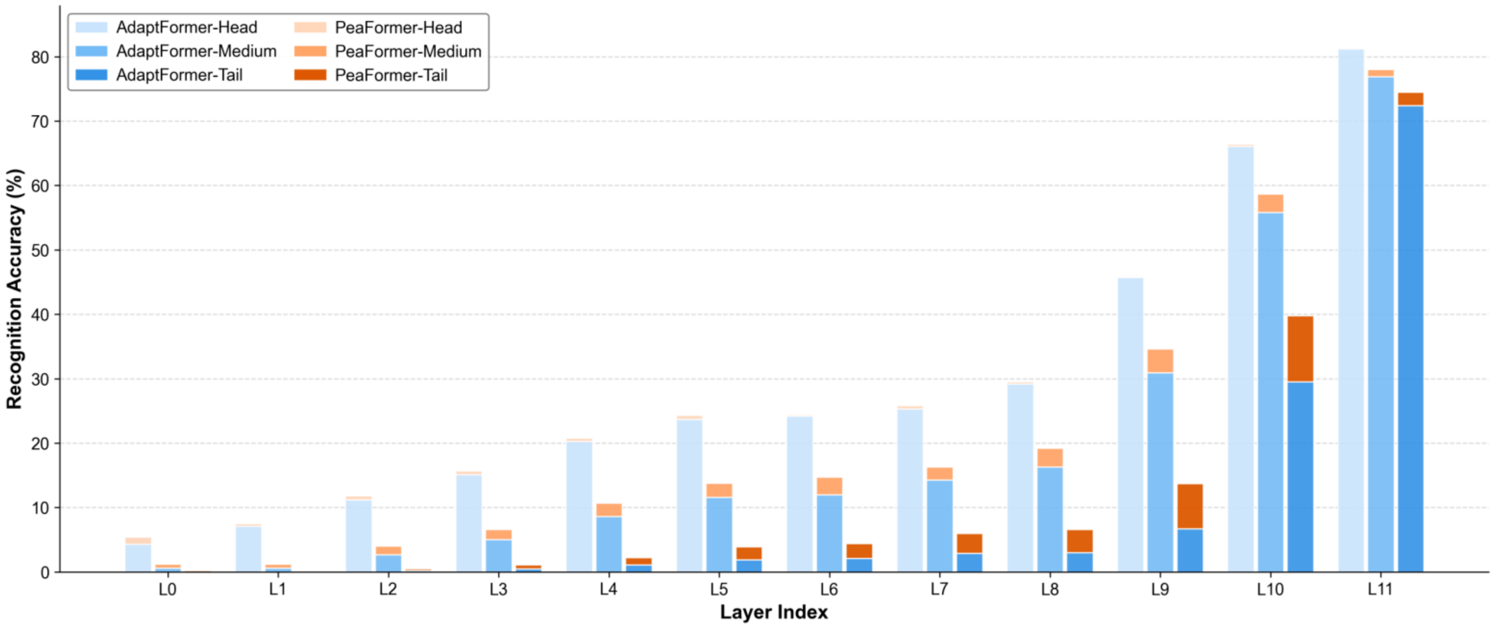
Scribes	Training set (12595)	Test set (3150)
Yin Zhi	3412	852
Guan Zhong	1508	377
Huang Men	4282	1071
Shi Fa	1670	420
Bao Xun	164	40
Cheng Wu	214	53
Chu Ju	445	111
Liang Chen	285	71
Bie Gua	32	8
Feng Xu Zhi Ming	161	40
Ming Xun	422	107



ResNet 34 Siamese model
Scribes Liang Chen and Bao Xu are more easily detected.
Scribes Feng Xu Zhi Ming and Ming Xun are detected more difficulty.

Panelist Position: Raw and Need for Synthetic Data

- **Minority classes**



Dealing with minority classes reduces classifier bias

Classification accuracy of Parameter-Efficient and Adaptive foundation model fine-tuning module for long-tailed ancient character recognition (PeaFormer) and Adapting vision transformers for scalable visual Recognition (AdaptFormer) across Head, Medium, and Tail classes at each layer of the CLIP image encoder, evaluated on the ImageNet-LT dataset.

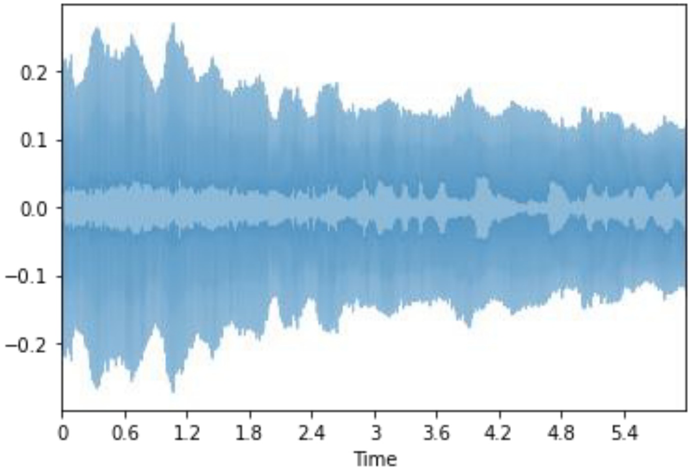
IARIA

Panelist Position: Raw and Need for Synthetic Data

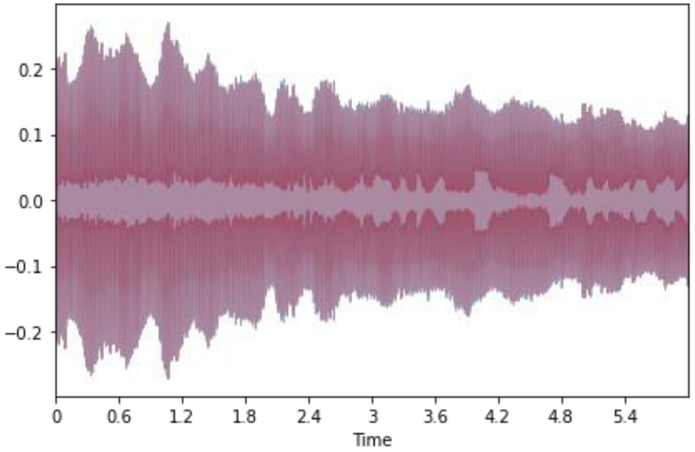
VALENCIA
March 2026

Pathological Speech Detection on Mobile Parkinson Disease Study (mPower)

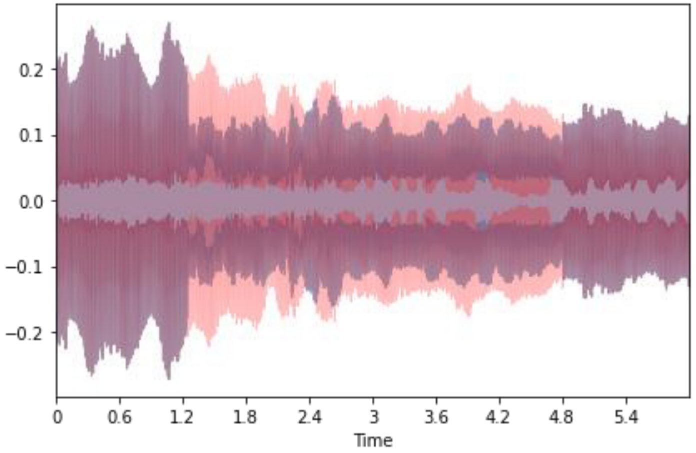
Augmentation (nplaug library of python)



Raw Parkinson speech



Adding noise

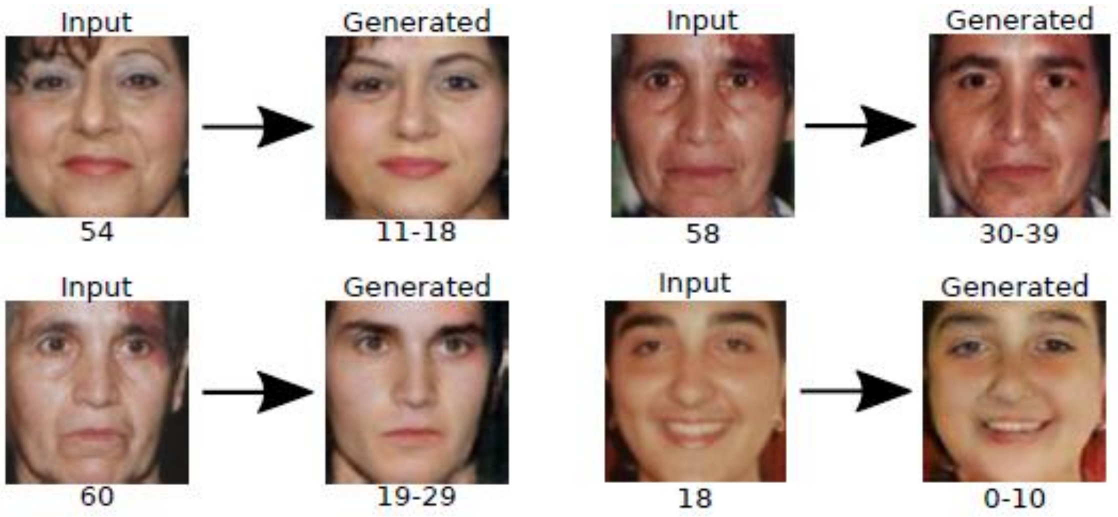


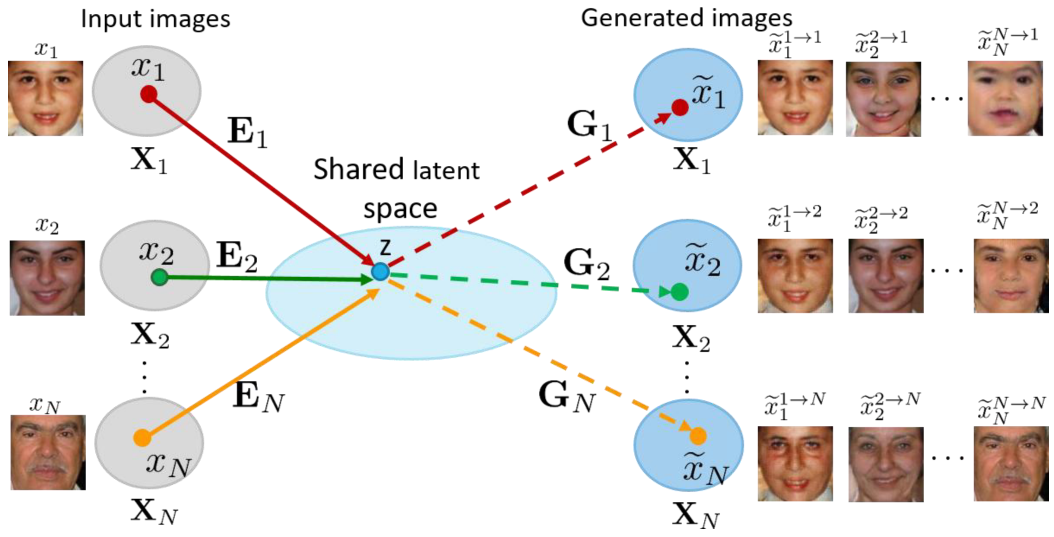
Pitch augmentation

DL-based algorithms are data hungry, but the raw data are never enough.

Generative Adversarial Networks (GANs)

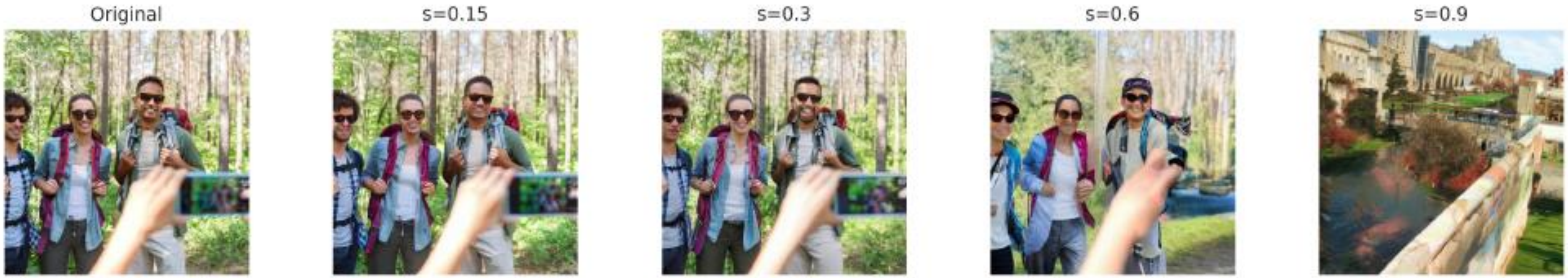
Face aging treated as an unsupervised image-to-image translation problem between face images of different age classes.





E. Pantraki and C. Kotropoulos, "Face aging using global and pyramid generative adversarial networks," Machine Vision and Applications 32, 82 (2021). https://doi.org/10.1007/s00138-021-01207-4

AI-Generated Images via Diffusion



Human-captured example (hikers group photo). Authentic photographs exhibit strong off-manifold divergence at higher noise strengths— fine details and spatial coherence collapse rapidly beyond s = 0.6.

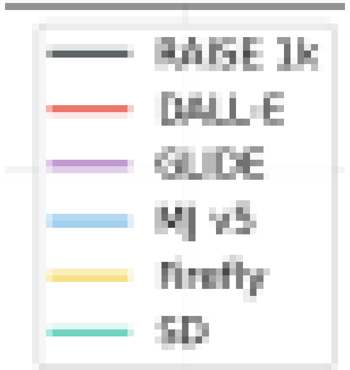
M. R. Ameen and A. Islam, Detecting AI-Generated Images via Diffusion Snap-Back Reconstruction: A Forensic Approach, https://arxiv.org/abs/2511.00352v1

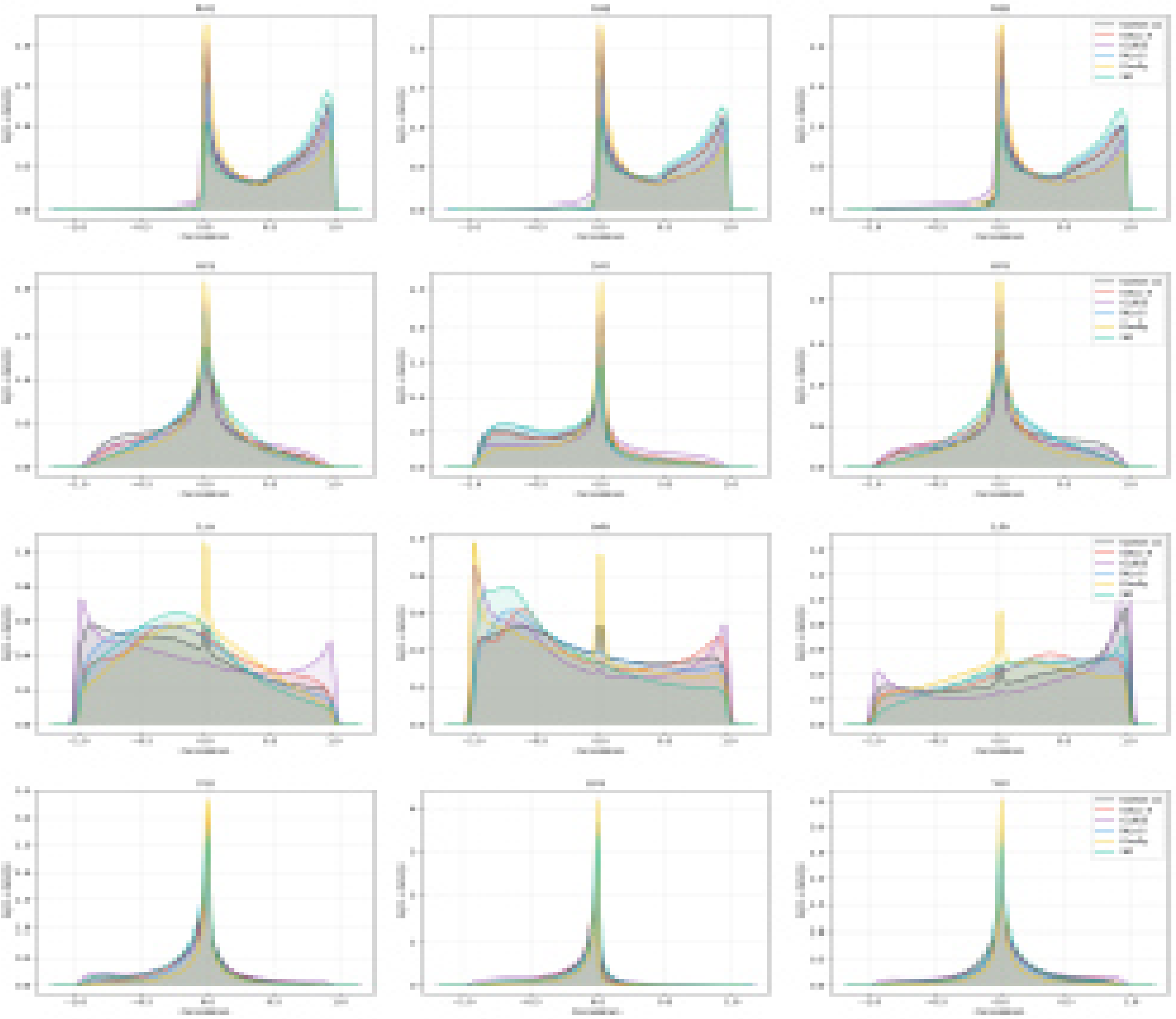
Panelist Position: Raw and Need for Synthetic Data

VALENCIA
March 2026

Detecting AI-Generated Images through Inter-Channel Color-Space Correlations

Log-density estimates of pairwise inter-channel correlation features across color spaces (RGB, HSV, Lab, YUV), comparing real images from RAISE-1k to synthetic images generated by DALL-E, GLIDE, Midjourney v5, Firefly, and Stable Diffusion from Synthbuster. Each row corresponds to one color space and shows the three channel-pair correlations.





Juan Pablo Sotelo et al.

Any

questions

or

comments?