20e édition

infoprodigital

Forum TERATEC 25

21 & 22 MAI 2025 | Parc Floral, Paris

LE RDV des experts du supercomputing

HPC – IA – Simulation – Quantique

En partenariat avec

L'USINENOUVELLE L'USINEDIGITALE IT INDUSTRIE & TECHNOLOGIES Teratec

Forum TERATEC

uni.lu | SnT

FROM SPACE TO THE CLASSROOM

DATA CHALLENGES AND LESSONS IN MULTIDISCIPLINARY RESEARCH PROJECTS

Jose Sosa

Forum TERATEC

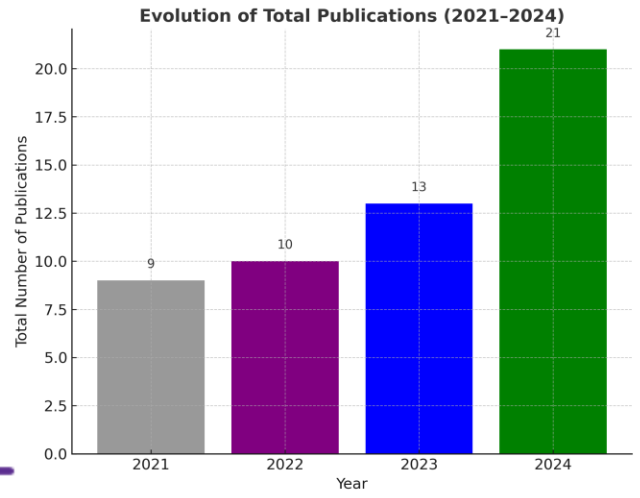
INTRODUCTION

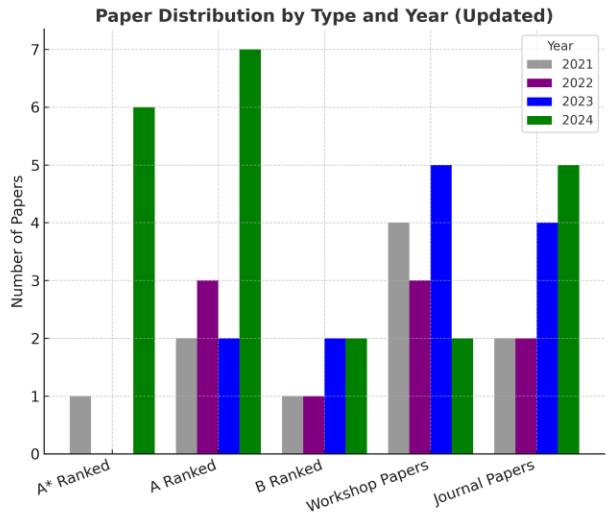
COMPUTER VISION, IMAGING AND MACHINE INTELLIGENCE RESEARCH GROUP (CVI2)

We conduct research in real-world applications of computer vision, image analysis, and machine intelligence, with extensive development of AI approaches.

23 researchers

NUMBER OF PUBLICATIONS PER YEAR





DATA

INTRODUCTION

RESEARCH DATA (in particular, for training DL systems) often differs from standard data.

- Dataset size
- File number and sizes
- Model size

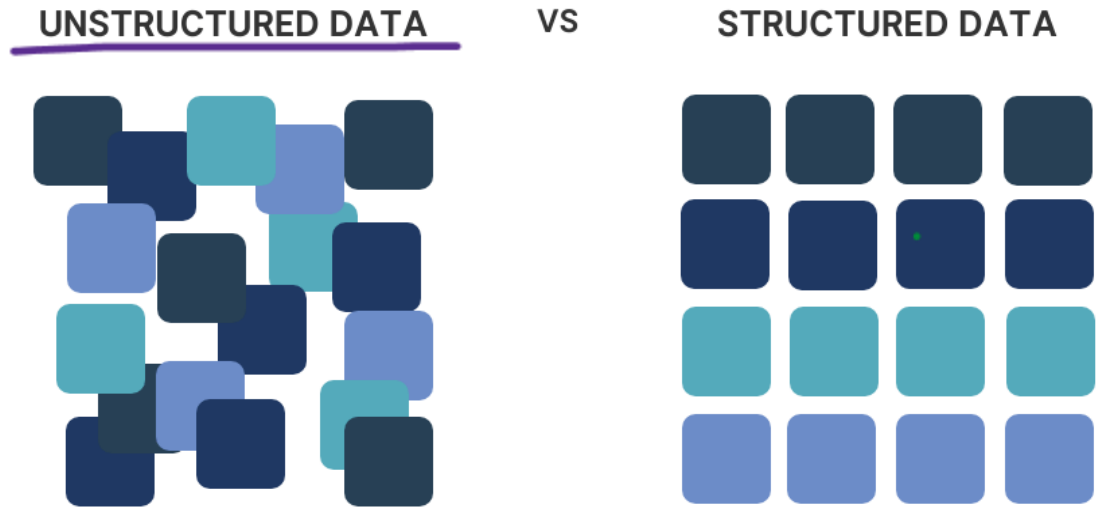


Image taken from: https://ubiai.tools/what-is-unstructured-data-what-issues-can-bring-to-dl-models/

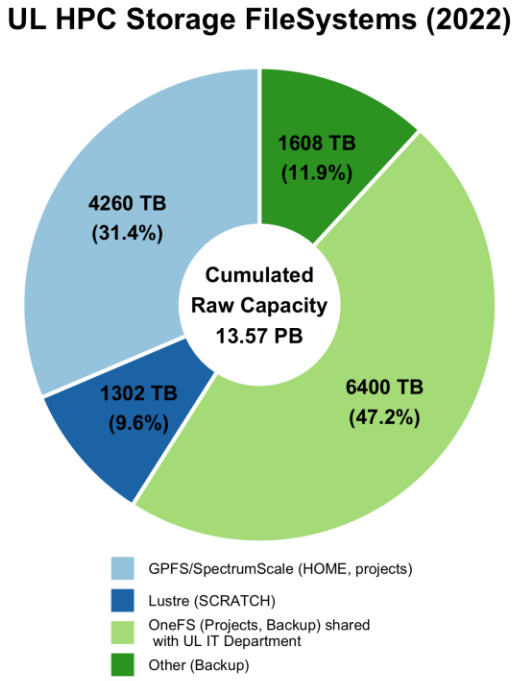
Forum TERATEC

INTRODUCTION

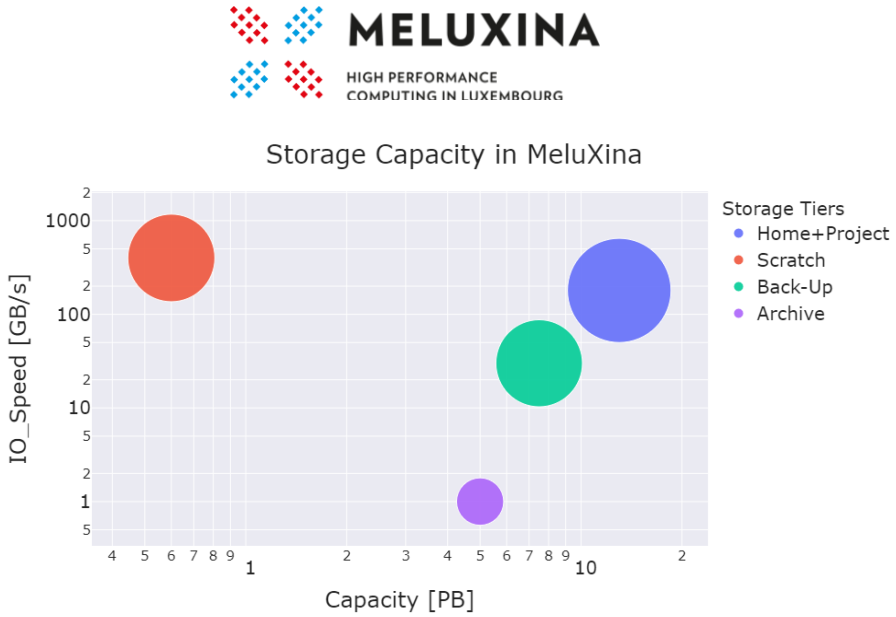
DATA STORAGE OPTIONS

6 servers, connected to UL shared storage

Lab clusters



University clusters (iris and ion)[1]



Meluxina[2]

[1] https://hpc-docs.uni.lu/systems/
[2] https://docs.lxp.lu/

Forum TERATEC

INTRODUCTION

Storage infrastructure is often a bottleneck, especially in fragmented systems with different access, latency, and format needs.

Data Heterogeneity: Structured vs unstructured.
Scalability: Need for scalable I/O pipelines for large-scale training
Access Patterns: Real-time querying vs batch pre-processing
Versioning: ML datasets and simulations evolve over time
Fragmentation: Tooling and storage formats vary widely by domain

Forum TERATEC

DOMAINS OF INTEREST FOR CVI2

1. Computer Vision
2. 3D CAD Design
3. Space Simulations
4. Education and Finance

DOMAINS OF INTEREST FOR CVI2: COMPUTER VISION

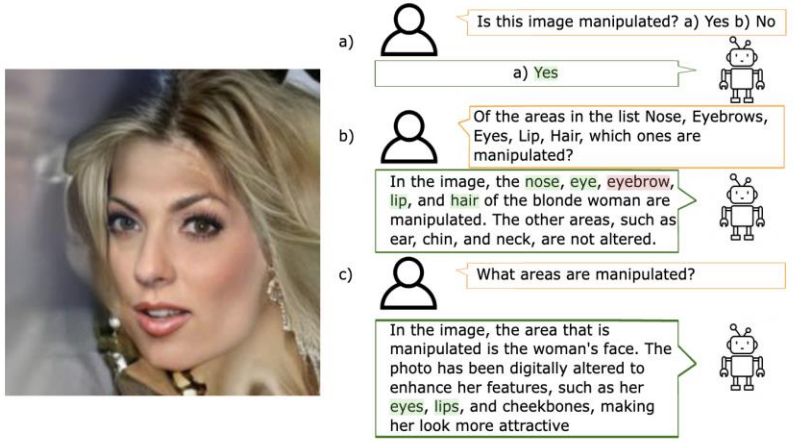
DEEPFAKE DETECTION

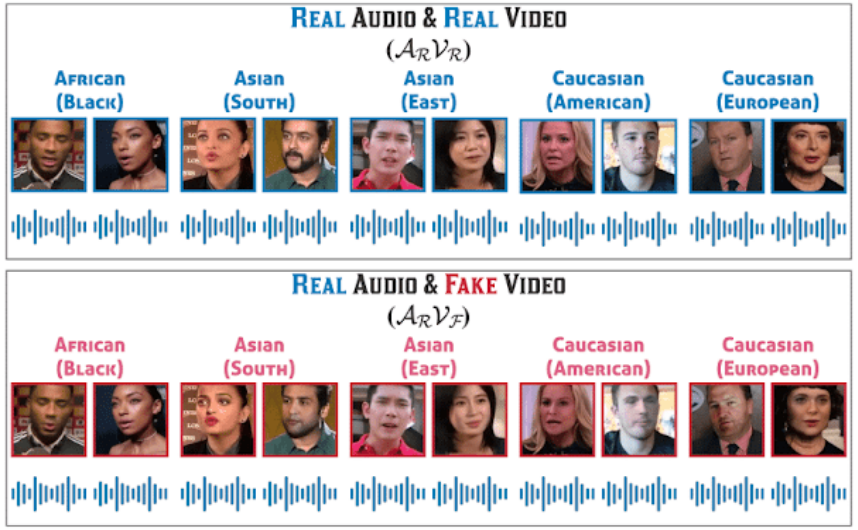
DATA:

- Videos
  - Image
  - Audio
- Text descriptions

CHALLENGE: Fast access to frame-level data and temporal coherence.

SOLUTIONS: Fast access data format for storage, e.g. HD5 files.





Astrid, M., Ghorbel, E., & Aouada, D. (2025). Audio-visual Deepfake Detection With Local Temporal Inconsistencies. arXiv preprint arXiv:2501.08137.

Foteinopoulou, N. M., Ghorbel, E., & Aouada, D. (2024). A Hitchhiker's Guide to Fine-Grained Face Forgery Detection Using Common Sense Reasoning. Advances in Neural Information Processing Systems, 37, 2943-2976.

Nguyen, D., Astrid, M., Ghorbel, E., & Aouada, D. (2024). FakeFormer: Efficient Vulnerability-Driven Transformers for Generalisable Deepfake Detection. arXiv preprint arXiv:2410.21964.

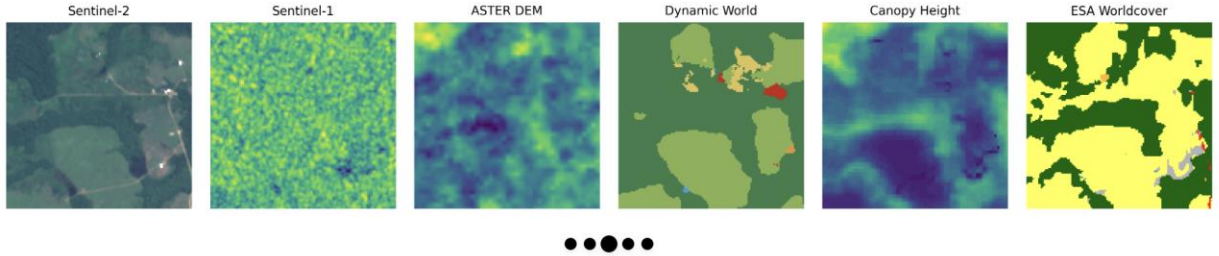
Nguyen, D., Astrid, M., Kacem, A., Ghorbel, E., & Aouada, D. (2025). Vulnerability-Aware Spatio-Temporal Learning for Generalizable and Interpretable Deepfake Video Detection. arXiv preprint arXiv:2501.01184.

DOMAINS OF INTEREST FOR CVI2: COMPUTER VISION

Date Sentinel-2: 2018-04-23
Geolocation: -7.7, -72.7
Biome: Tropical & Subtropical
Moist Broadleaf Forests
Ecoregion: Iquitos várzea

Temperature [C]
yearly mean/min/max: 25.4/15.1/39.8
monthly mean/min/max: 24.7/21.5/31.9

Precipitation [mm]
yearly total: 8815.1
monthly total: 204.7



REMOTE SENSING

DATA:

- RGB
- S2 images
- S1 images
- Radar data
- Text descriptions
- Numeric data

CHALLENGE: Efficient storage and access of multi modal data (600GB).

SOLUTIONS: Efficient storage, e.g. HD5 files.

Sosa, J., Aloulou, M., Rukhovich, D., Sleimi, R., Changaival, B., Kacem, A., & Aouada, D. (2024). How Effective is Pre-training of Large Masked Autoencoders for Downstream Earth Observation Tasks?.
Sosa, J., Rukhovich, D., Kacem, A., & Aouada, D. (2025). MultiMAE Meets Earth Observation: Pre-training Multi-modal Multi-task Masked Autoencoders for Earth Observation Tasks.

Forum TERATEC

DOMAINS OF INTEREST FOR CVI2: 3D CAD DESIGN

3D CAD DESIGN

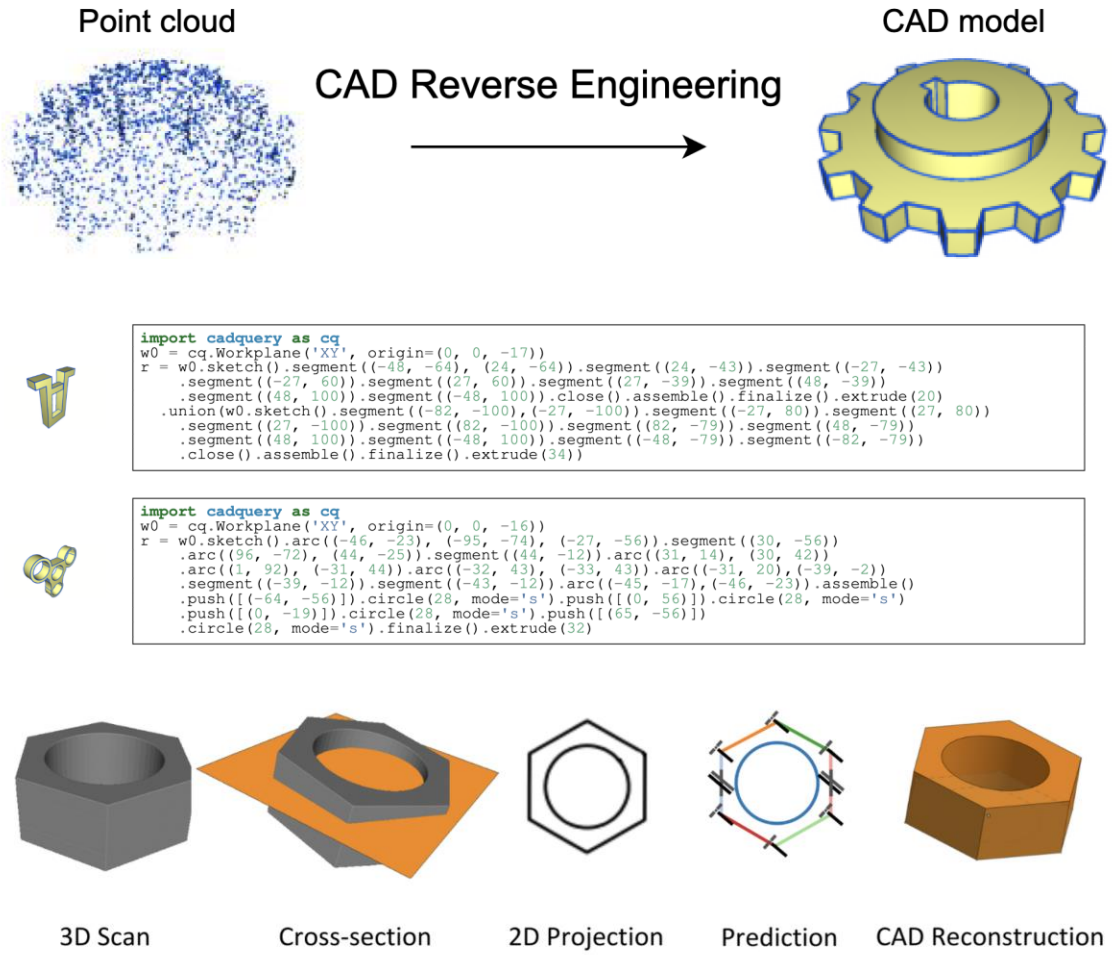
DATA:

- Point clouds
- Code
- Text
- 3D scans
- Sketches

CHALLENGE: Multi-format input, CAD versioning, human-in-the-loop annotations, and data sharing.

SOLUTIONS: Reduce storage by generatering representations on the fligth.

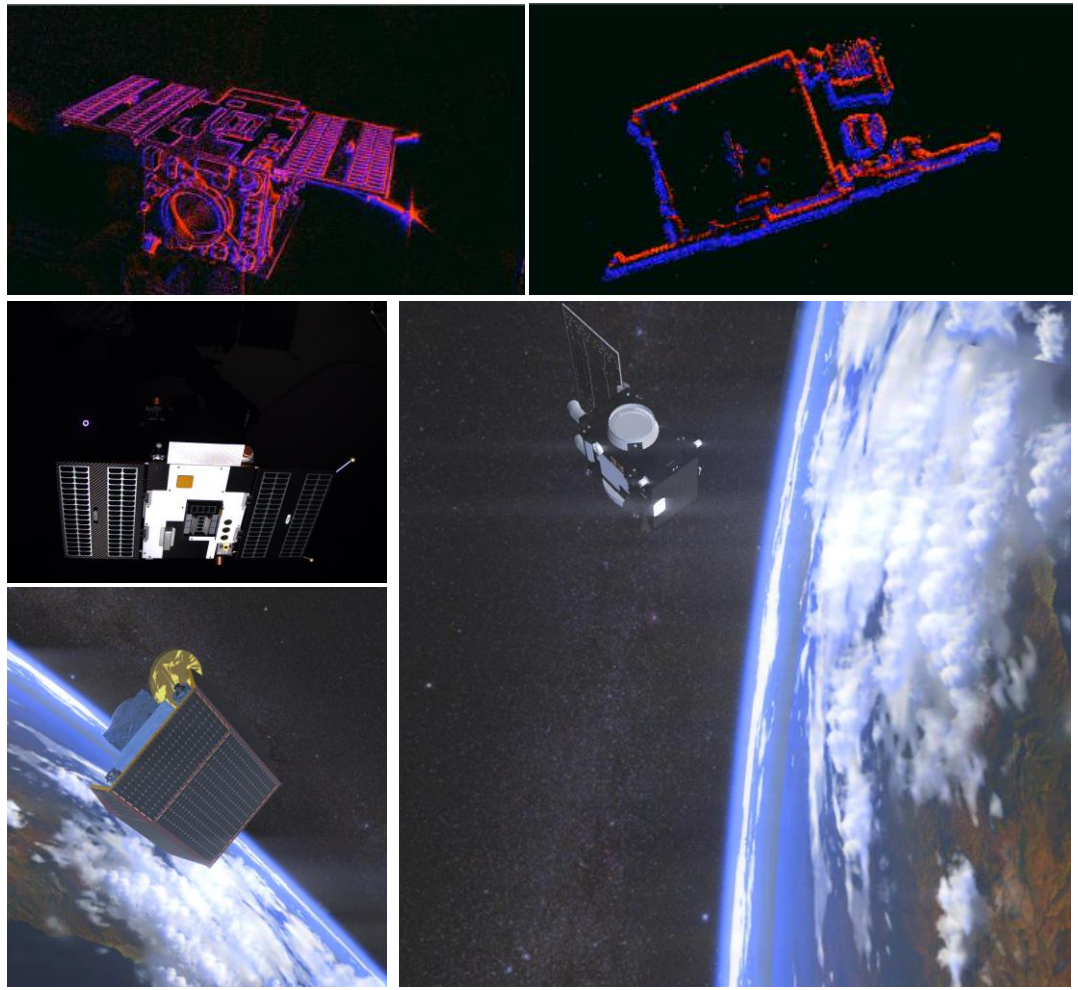
Demo: https://huggingface.co/spaces/filapro



Rukhovich, D., Dupont, E., Mallis, D., Cherenkova, K., Kacem, A., & Aouada, D. (2024). CAD-Recode: Reverse Engineering CAD Code from Point Clouds. arXiv preprint arXiv:2412.14042.

KARADENIZ, A. S., MALLIS, D., MEJRI, N., CHERENKOVA, K., KACEM, A., & AOUADA, D. (2024, October). DAVINCI: A Single-Stage Architecture for Constrained CAD Sketch Inference. In British Machine Vision Conference 2024.

DOMAINS OF INTEREST FOR CVI2: SPACE



SPACE SIMULATIONS

DATA:

- Events
- RGB
- Numeric data: trajectories, 6DoF poses.
- 3D models

CHALLENGE: High-resolution images, physics simulation logs, mesh/texture pairs, and data sharing.

Rathinam, A., Qadadri, H., & Aouada, D. (2024, May). Spades: A realistic spacecraft pose estimation dataset using event sensing. In 2024 IEEE International Conference on Robotics and Automation (ICRA) (pp. 11760-11766). IEEE.

Rathinam, A., Ali, M. M., Gaudilliere, V., & Aouada, D. (2024). SPARK 2024: Datasets for Spacecraft Semantic Segmentation and Spacecraft Trajectory Estimation. Feb, 2, 7.

DOMAINS OF INTEREST FOR CVI2: EDUCATION AND FINANCE

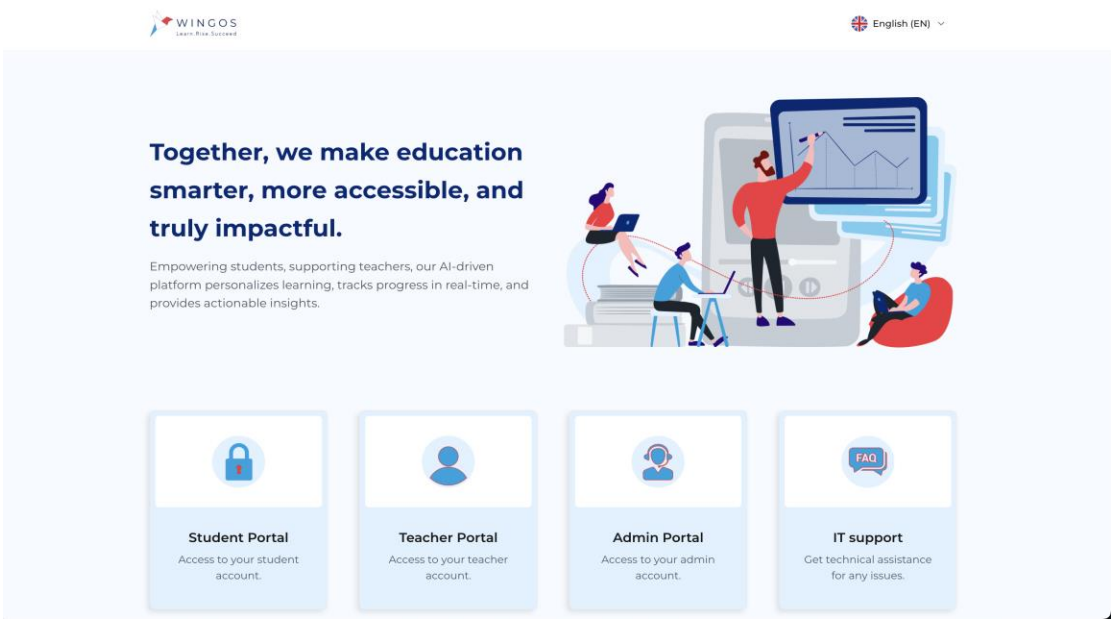
LLMS IN CLASSROOMS AND FINANCE

DATA:

- Logs
- Text
- Prompts
- Interactions

CHALLENGE: Fine-grained access control, logging, versioned corpora.

SOLUTIONS: Lightweight document stores + secure audit trails.



Forum TERATEC

LESSONS LEARNED

1. Anticipate how your data will be used
2. Know your use case
3. Keep raw data raw
4. Store data in open formats
5. Data should be structured for analysis
6. Data should be uniquely identifiable
7. Link relevant metadata
8. Adopt the proper privacy protocols
9. Have a systematic backup scheme
10. The location and method of data storage depend on how much data you have

From: Hart, E. M., Barmby, P., LeBauer, D., Michonneau, F., Mount, S., Mulrooney, P., ... & Hollister, J. W. (2016). Ten simple rules for digital data storage. PLoS computational biology, 12(10), e1005097.

Forum TERATEC

uni.lu | SnT

THANK YOU
MERCI
GRACIAS
DANKE

LET'S CONNECT!!

Email: jose.sosa@uni.lu