Data (Quality) Challenges in Multimodal AI Pipelines

Laure Berti-Equille

IRD, ESPACE-DEV
Montpellier, France

https://laureberti.github.io/website/
laure.berti@ird.fr

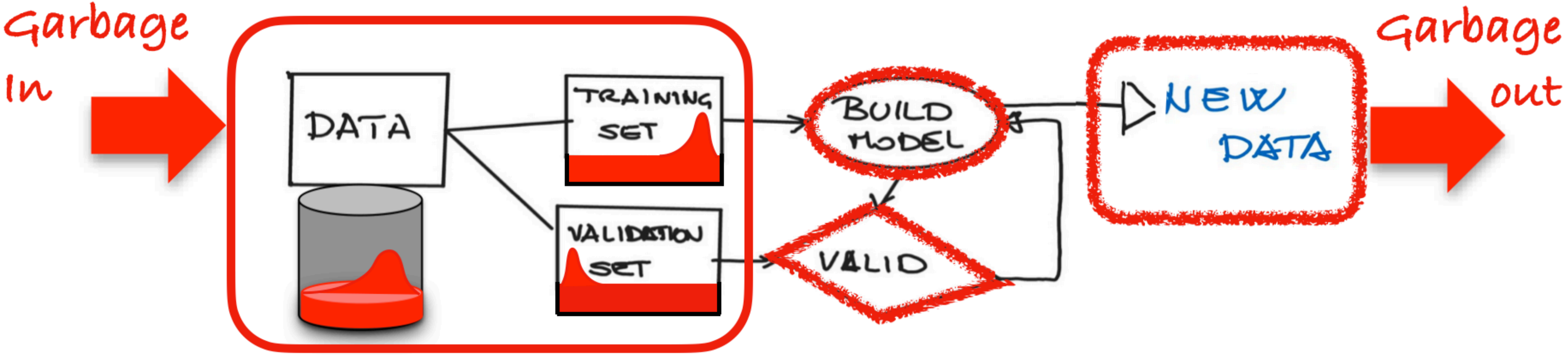
Institut de Recherche pour le Développement
FRANCE

April 4, 2025

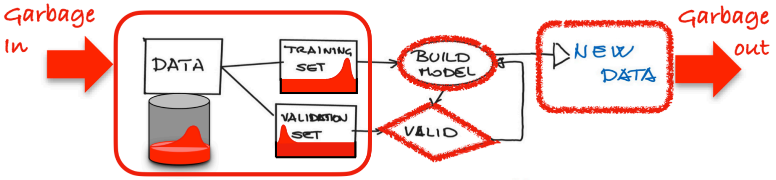
Outline

1. Motivations
2. Challenges in MML
3. Methods & Contributions
4. Conclusions

Motivations (1/4)



Motivations (1/4)



Data quality profiling is always required.

Relational data quality problems

Nobel Laureates in Chemistry

Misfielded Value

Representation

Name	Institution	Institution_City	DoB
Skłodowska-Curie Marie	Institut Pasteur	Varsovie	07-11-1867
M. Curie	Pasteur Institute	Paris	1867-11-07
Melvin Calvin	UC Berkeley	Berkeley	1911-04-08
Marie Curien	Paris	Pasteur Institute	2007-11-07
Avram Hershko	NULL	Haifa	NULL
Ronald Hoffman		US	00000000

Duplicates

Typos

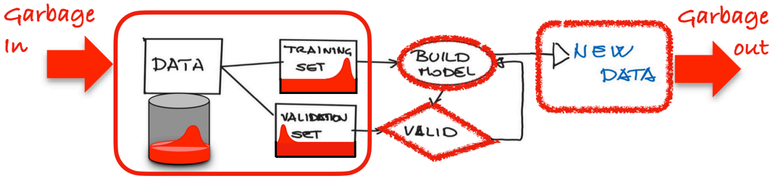
Inconsistencies

Incorrect Values

Incorrect Value

Missing Values

Motivations (1/4)



Data quality profiling is always required.

Relational data quality problems

Nobel Laureates in Chemistry

Misfielded Value

Representation

Name	Institution	Institution_City	DoB
Skłodowska-Curie Marie	Institut Pasteur	Varsovie	07-11-1867
M. Curie	Pasteur Institute	Paris	1867-11-07
Melvin Calvin	UC Berkeley	Berkeley	1911-04-08
Marie Curien	Paris	Pasteur Institute	2007-11-07
Avram Hershko	NULL	Haifa	NULL
Ronald Hoffman		US	00000000

Duplicates

Typos

Inconsistencies

Incorrect Value

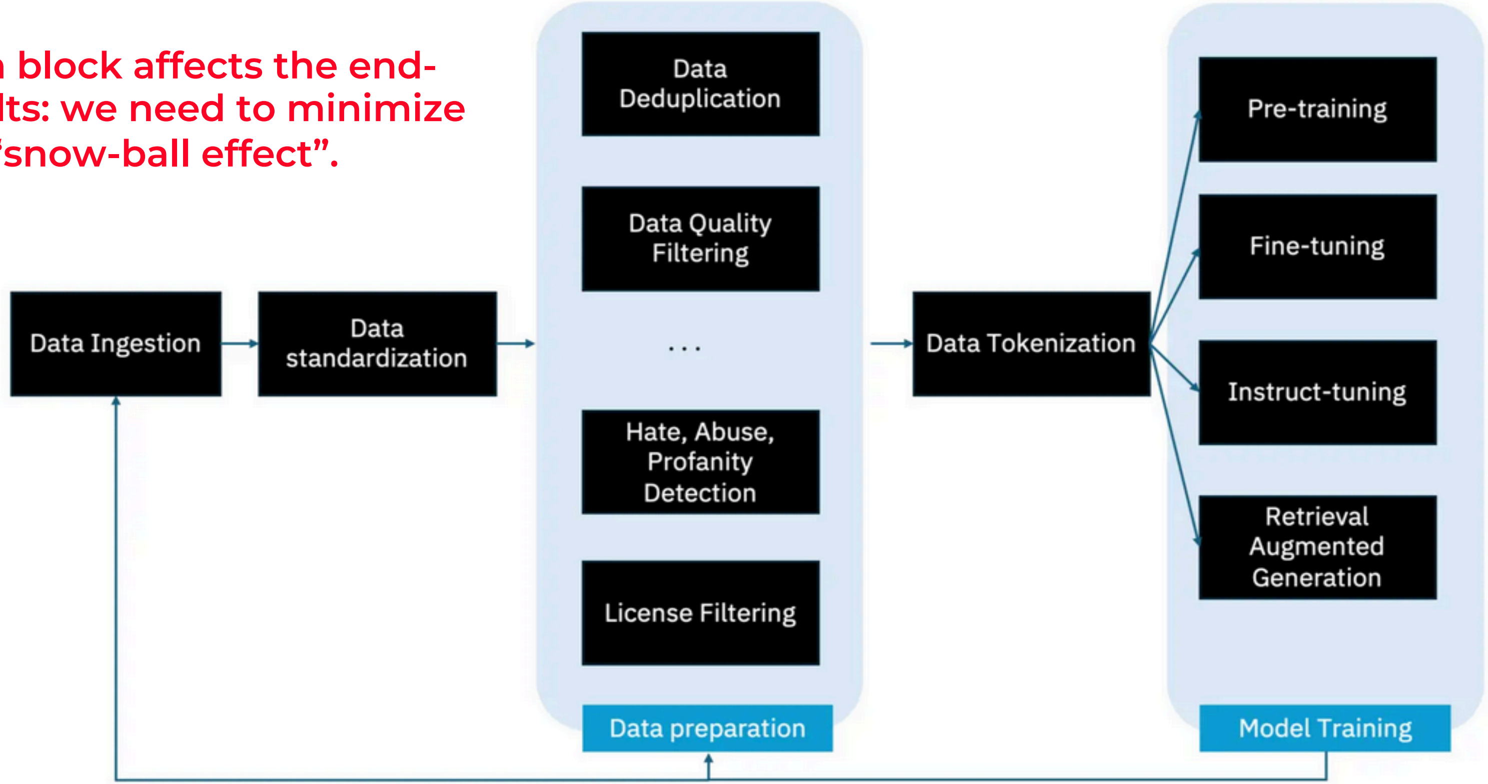
Incorrect Values

Missing Values

Conflicts intra-/ intermodality

Motivations (2/4): Data-centric ML pipeline

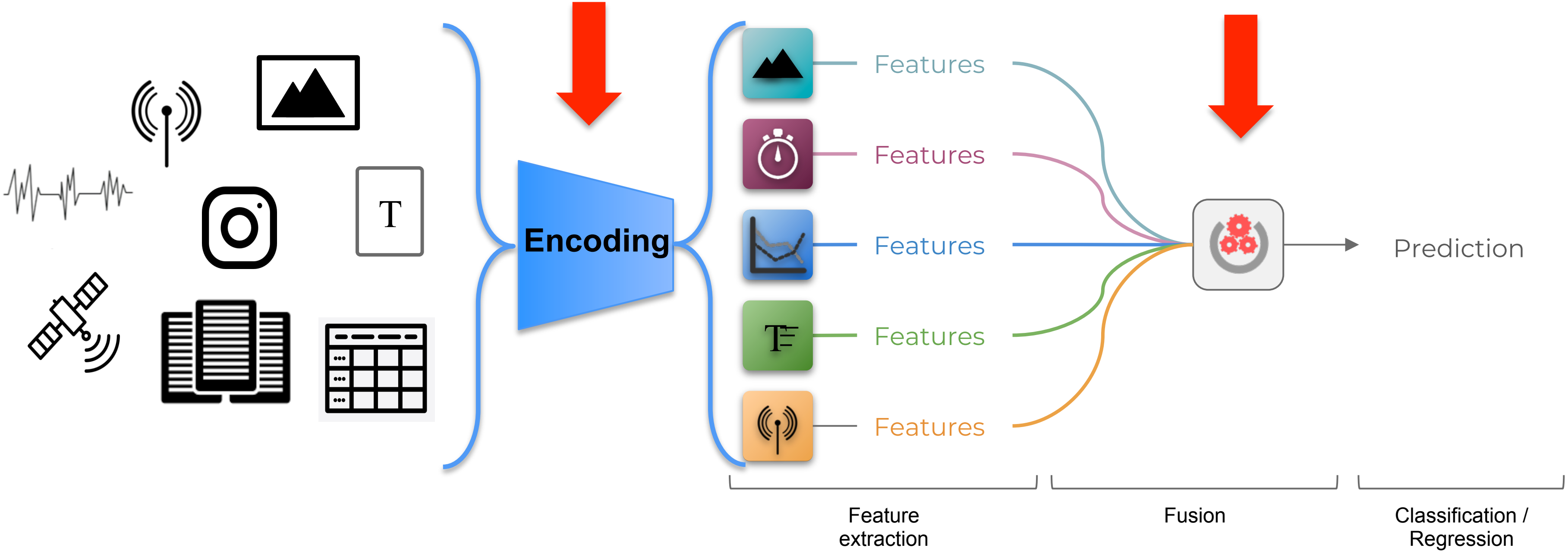
Each block affects the end-results: we need to minimize the "snow-ball effect".



From Hima Patel et al., https://fr.slideshare.net/slideshow/data_prep_techniques_challenges_methods-pdf-a190/271527890

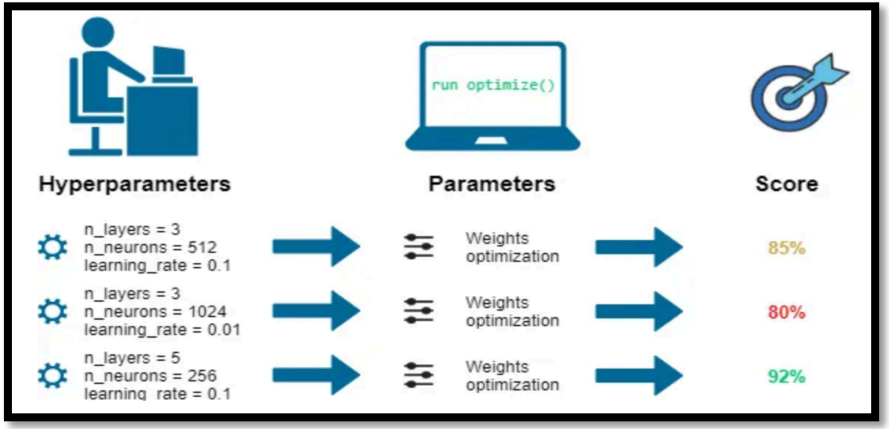
Motivations (3/4): Multimodal Learning

We need to select the optimal encoding and fusion functions



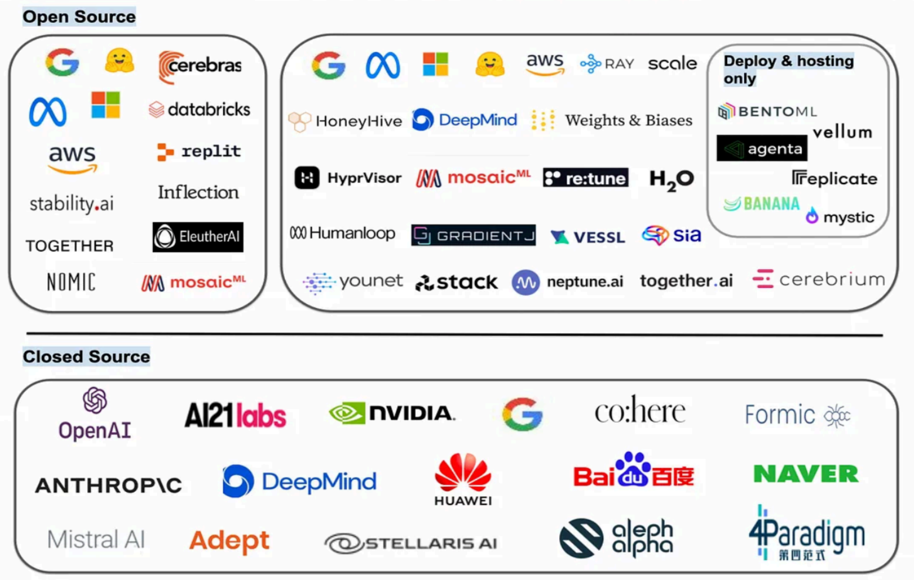
Motivations (4/4): Reproducibility & Traceability

- Ensure stable and consistent hyperparameter optimization



- Ensure resilience to multimodal data poisoning attacks

- Trace back pre-training, fine-tuning, and prompt engineering



We need reproducibility

We need traceability and explainability

Theoretical, Technical, and Experimental Challenges

Multimodal Deep Learning

- Complex models → MM Network Architecture Search
- Costly training → Cost-driven and frugal models
- Hard to communicate to non-experts → Beyond XAI: Chain-of-evidences

(Multimodal) Uncertainty Quantification

- Quantify aleatoric and epistemic uncertainty
- Detect multimodal contradictions

image contradiction

Text/image contradiction

Wish list Before Integrating/Using LLMs & MLLMs

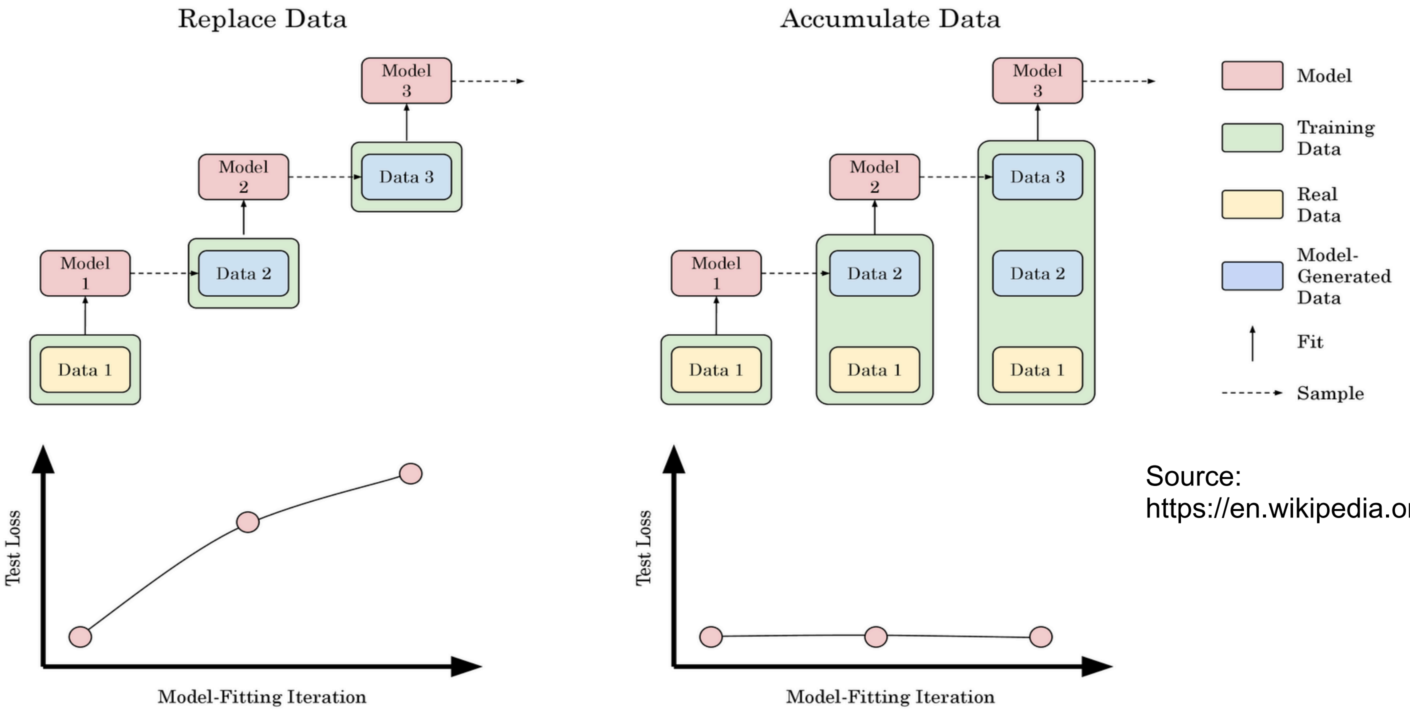
We need to:

- Quantify **LLM hallucination & factuality** in perspective with the model/ training size
- Detect **stereotype amplification** due to bias and low quality training corpus
- Evaluate **sensitivity** to prompt variations, noise, conflicting (multimodal) data or domain shift
- Evaluate LLM **vulnerability to adversarial attacks** (e.g., generated texts used in pretraining or prompts)
- Use dedicated **benchmarks** and design **controlled experiments**

Emily Bender, Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell, "On the Dangers of Stochastic Parrots: Can Language Models Be Too Big? 🦜" In Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency, pp. 610-623. 2021.

The Understudied Model Collapse Phenomenon

Increasing use and re-use of LLM-generated data and synthetic data



Source:
https://en.wikipedia.org/wiki/Model_collapse

Shumailov, I., Shumaylov, Z., Zhao, Y. *et al.* AI models collapse when trained on recursively generated data. *Nature* **631**, 755–759 (2024). https://doi.org/10.1038/s41586-024-07566-y

Inadequacy/Inexistence of benchmarks: e.g. MM Fact-checking

Name	# Claims	# Labels	Data	Year
LIAR [4]	12836	6	Claim Text, Metadata (Speaker etc.)	2017
CREDBANK [8]	1049	5	Claim Text, Event, Topic	2015
The Lie Detector [9]	600	2	Claim Text	2009
Claim matching beyond english [10]	2343	3	Claim Text Pairs	2021
FEVER [1]	185445	3	Claim Text, Document Text	2018
MultiFC [12]	36534	40	Claim Text, Document url, Metadata	2019
Fakeddit [13]	1 million	2/3/6	Claim Text, Claim image	2019
Covid-19 Fake News dataset [11]	10700	2	Claim Text	2020
FakeNewsNet [14]	23921	2	Claim Text, Spatiotemporal info	2019
Whatsapp fact-checking dataset [15]	1032	3	Claim Image, Metadata	2020
Factify (ours)	50000	5	Claim Text, Claim Image, Document Text, Document Image, Images OCR	2021

Table 1
Details of related public datasets for automated fact-checking along with available meta data and release year.

Mishra et al. FACTIFY: A Multi-Modal Fact Verification Dataset, De-Factify: Workshop on Multimodal Fact Checking and Hate Speech Detection, co-located with AAAI 2022. https://ceur-ws.org/Vol-3199/paper18.pdf

Our contributions

Ph.D. thesis of Grigor Bezirganyan, and collaborators @AMU

1. **M2-Mixer**: Design adaptive, conceptually, computationally simple, scalable multimodal deep learning architecture

2. **MixMax:** Find the optimal multimodal deep learning architecture

3. **DBF:** Quantify the uncertainties in multimodal learning

4. **LUMA:** Provide a benchmark dataset for multimodal learning and uncertainty quantification

Multimodal Data Fusion - Main Contributions

Contribution 1:
Propose an all MLP-based approach for multimodal fusion

Contribution 2:
Improve modality representations by optimizing the learning process with multi-head loss

Contribution 3:
Propose a micro-benchmarking pipeline for automatic MLP-based multimodal architecture design

Multimodal Data Fusion - Related Work

Current state-of-the-art model mainly use

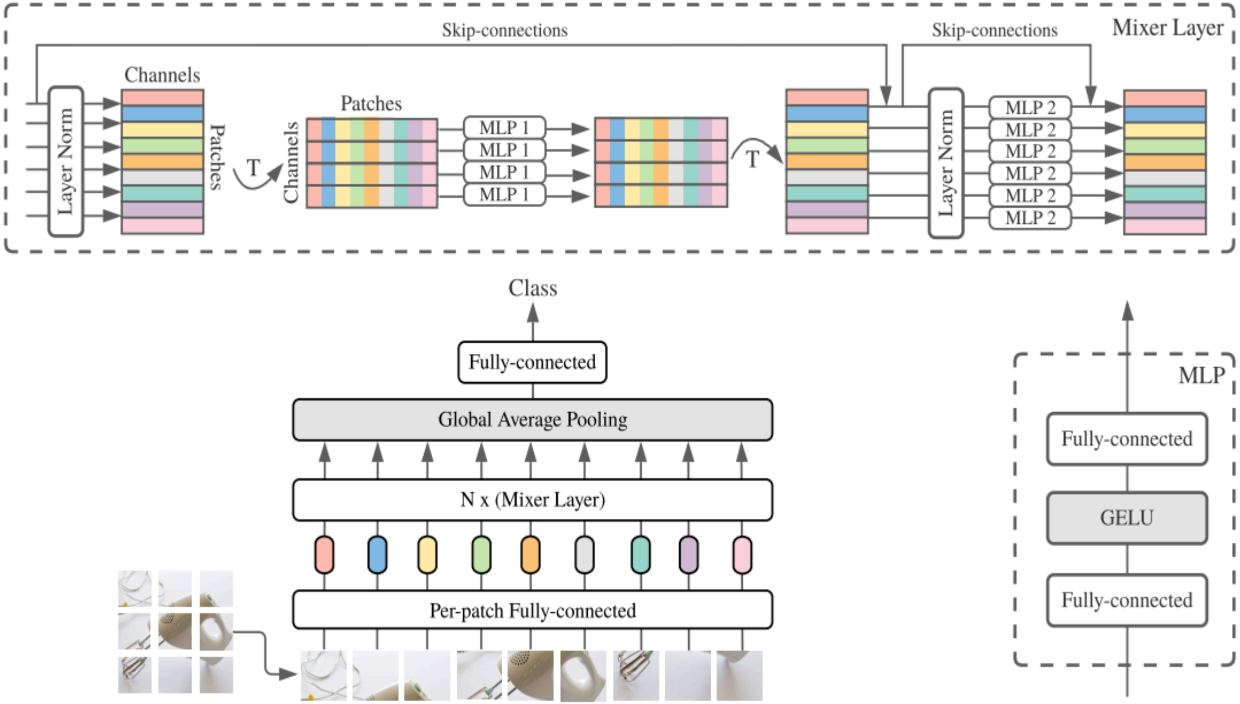
- Big Convolutional Networks
- Transformers
- Neural Architecture Search
- Pre-trained models
- Complex fusion functions

These approaches are often conceptually, computationally complex

Multimodal Networks may favor one modality over the other, and find suboptimal representations for the modalities [Wang et al., 2020]

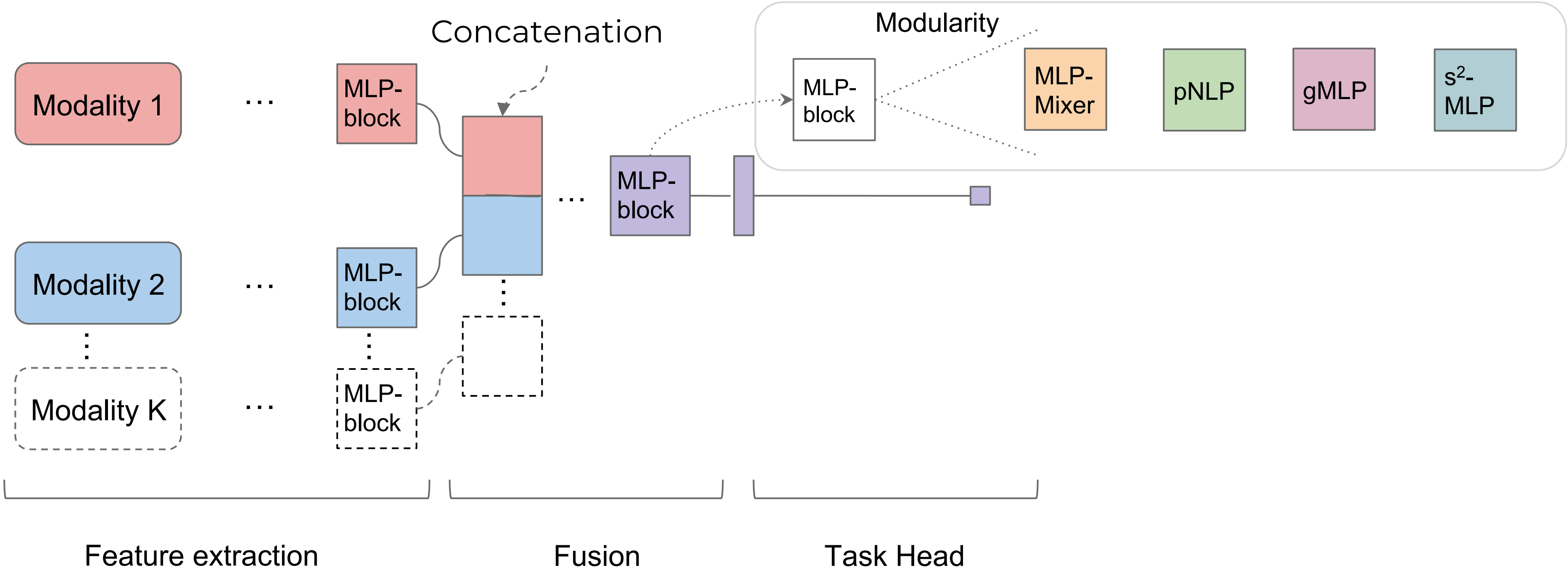
Multimodal Data Fusion - Related Work - MLP Mixers

- Only MLPs
- Conceptually / Computationally Simple
- Competitive with Transformers / CNNs



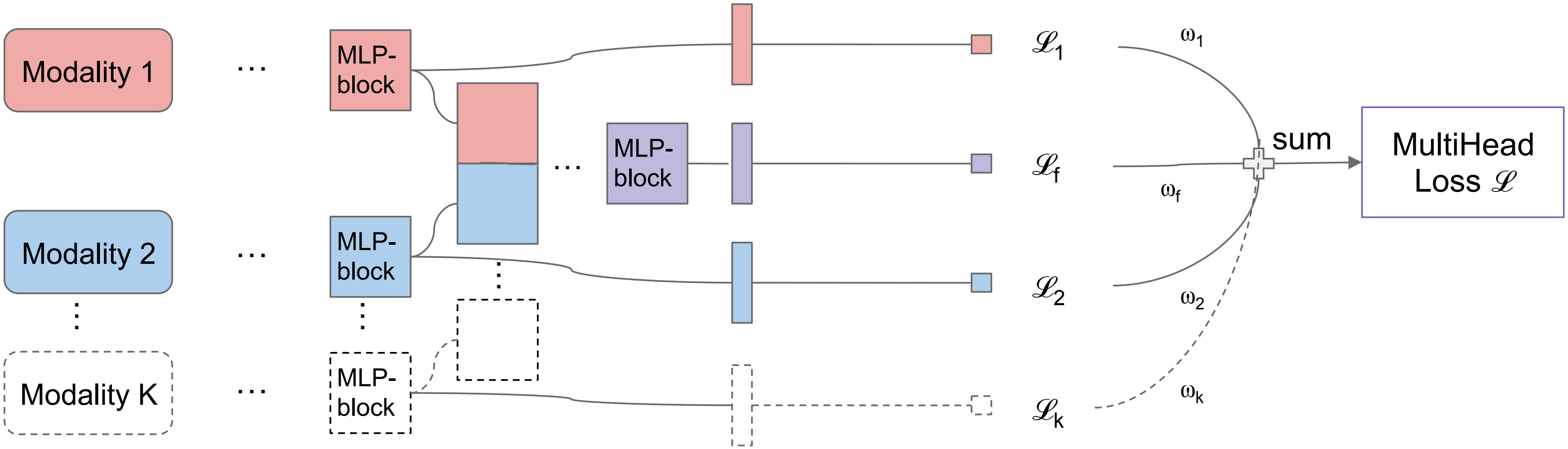
(Tolstikhin et. al, 2021)

Multimodal Data Fusion - Multimodal Mixer



G. Bezirganyan, S. Sellami, L. Berti-ÉQuille and S. Fournier, "M2-Mixer: A Multimodal Mixer with Multi-head Loss for Classification from Multimodal Data," 2023 IEEE International Conference on Big Data (BigData), Sorrento, Italy, 2023, pp. 1052-1058, doi: 10.1109/BigData59044.2023.10386252.

Multimodal Data Fusion - M2-Mixer



$$\mathcal{L}(\hat{y}_f, \hat{y}_1, \hat{y}_2, y) = w_f \mathcal{L}_f(\hat{y}_f, y) + w_1 \mathcal{L}_1(\hat{y}_1, y) + w_2 \mathcal{L}_2(\hat{y}_2, y) + \cdots + w_k \mathcal{L}_k(\hat{y}_k, y)$$

G. Bezirganyan, S. Sellami, L. Berti-ÉQuille and S. Fournier, "M2-Mixer: A Multimodal Mixer with Multi-head Loss for Classification from Multimodal Data," 2023 IEEE International Conference on Big Data (BigData), Sorrento, Italy, 2023, pp. 1052-1058, doi: 10.1109/BigData59044.2023.10386252.

Multimodal Data Fusion - Experiments

2 Datasets:

	AV-MNIST [Vielzeuf et al., 2018]	MIMIC-III [Johnson et al., 2015]
Field	Multimedia	Healthcare
Modalities	Image / Audio	Time Series / Tabular
Samples train / val / test	55,000 / 5000 / 10000	26,093 / 3,261 / 3,261

Our models:

- MMixer (no multi-head loss)
- M2-Mixer (with multi-head loss)

9 Baseline models:

Simple Late Fusion [Liang et al, 2021], LRTF [Liu et al., 2018], MFAS [Pérez-Rúa et al., 2019], RefNet [Sankaran et al., 2021], MVAE [Wu et al., 2018], MFM [Tsai et al., 2019], CCA [Sun et al., 2020], MI-Matrix [Jayakumar et al., 2020], GradBlend [Wang et al., 2020]

Multimodal Data Fusion - Results

AV-MNIST Image / Audio 55,000 / 5000 / 10000

Architecture	Accuracy % - avg (10 runs)	Accuracy % - max	Train time (s)
MFAS	72.64 ± 0.2	72.93	6,710 ± 12817
GradBlend	68.71 ± 0.7	69.51	43768 ± 5554
M2-Mixer B	**73.06 ± 0.2**	**73.34**	10271 ± 6578
M2-Mixer M	72.81 ± 0.2	73.20	**4147 ± 1642**

MIMIC-III Time Series / Tabular 26,093 / 3,261 / 3,261

Architecture	Accuracy % - avg (10 runs)	Accuracy % - max	Train time (s)
MFAS	78.02 ± 0.4	78.63	8043 ± 663
GradBlend	78.1 ± 0.3	78.51	7988 ± 239
M2-Mixer H	78.32 ± 0.3	**79.03**	840 ± 119
M2-Mixer LC	**78.43 ± 0.3**	78.76	**597 ± 113**

Our Proposed Models

B, M, H, LC are different configurations of the same model

Number of Parameters

B : 8.3 m
M : 88 k
H : 2.9 k
LC : 2.9K

M2-Mixer outperforms MFAS and GradBlend with much lower training time

G. Bezirganyan, S. Sellami, L. Berti-ÉQuille and S. Fournier, "M2-Mixer: A Multimodal Mixer with Multi-head Loss for Classification from Multimodal Data," 2023 IEEE International Conference on Big Data (BigData), Sorrento, Italy, 2023, pp. 1052-1058, doi: 10.1109/BigData59044.2023.10386252.

Our contributions

Ph.D. thesis of Grigor Bezirganyan, and collaborators @AMU

1. **M2-Mixer**: Design adaptive, conceptually, computationally simple, scalable multimodal deep learning architecture
2. **MixMax:** Find the optimal multimodal deep learning architecture
3. **DBF**: Quantify the uncertainties in multimodal learning
4. **LUMA:** Provide a benchmark dataset for multimodal learning and uncertainty quantification

Architecture search for M2-Mixers

M2-Mixer:

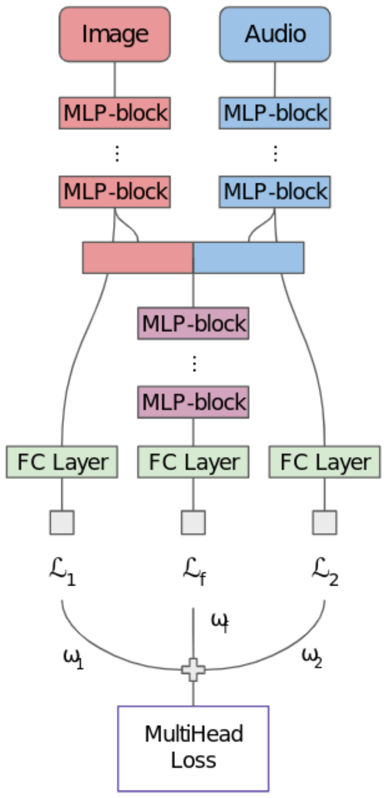
- Use MLP-blocks to extract information from each modality
- Use MLP-blocks for fusing the extracted features
- Use Multi-head loss for optimisation
- MLP-blocks can be any MLP-based architecture

Question:

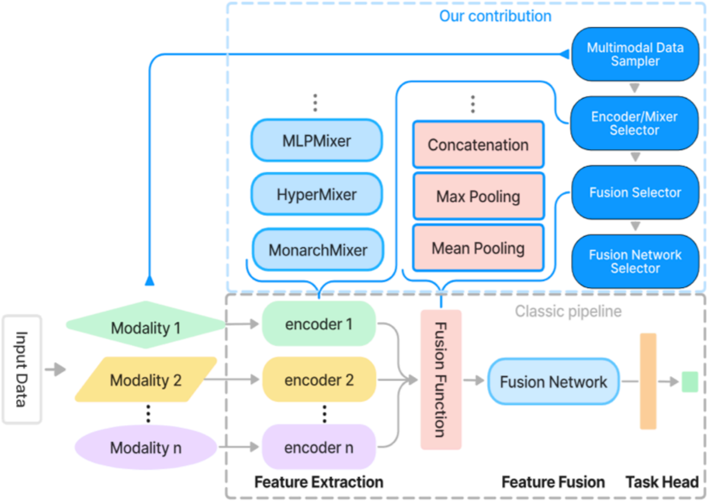
- What MLP-based architecture to use for each MLP-Block?

MLP Mixer | Hyper Mixer | Monarch Mixer | RaMLP | ...

M2-Mixer



Contributions:



Contribution 1:

Propose a flexible pipeline that:

- Takes a small sample of the dataset
- Conducts micro-benchmarking on the subset
- Constructs optimal MLP-based networks based on the micro-benchmarking

Contribution 2:

Experimentally validate that our pipeline enhances accuracy over standard MLP-based multimodal networks.

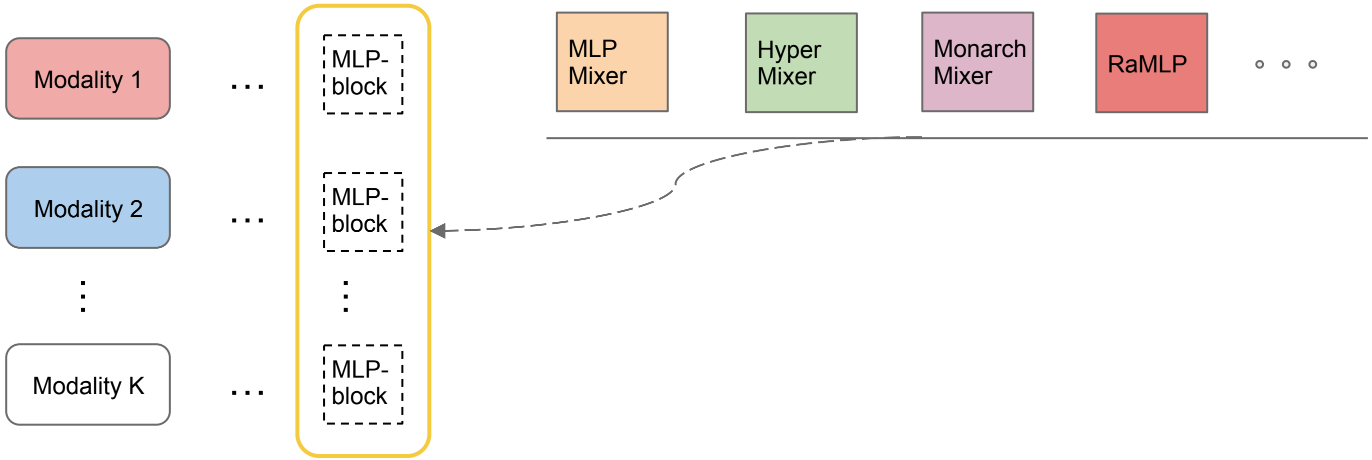
Abdelmadjid Chergui, Sana Sellami, Laure Berti-Équille, Sébastien Fournier. MixMAS: A Framework for Sampling-Based Mixer Architecture Search for Multimodal Fusion and Learning. 2024 IEEE International Conference on Big Data, MMAI 2024 workshop, Dec 2024, Washington DC, USA.

MixMAS: Sampling based micro-benchmarking

1. Take a representative small sample of the dataset [Hogg et al., 2023]

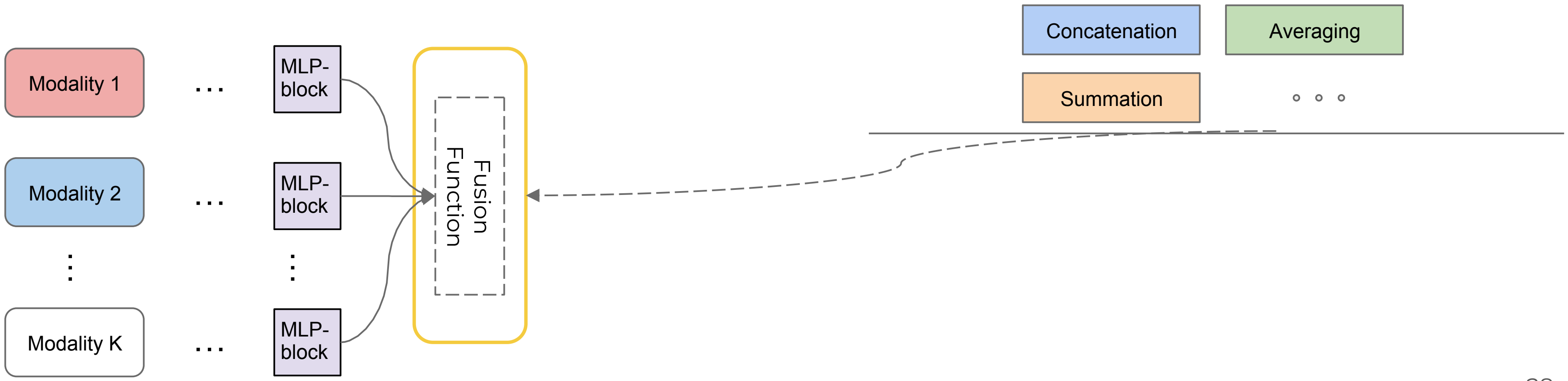
MixMAS: Sampling based micro-benchmarking

1. Take a representative small sample of the dataset [Hogg et al., 2023]
2. Find the best uni-modal encoders for each modality



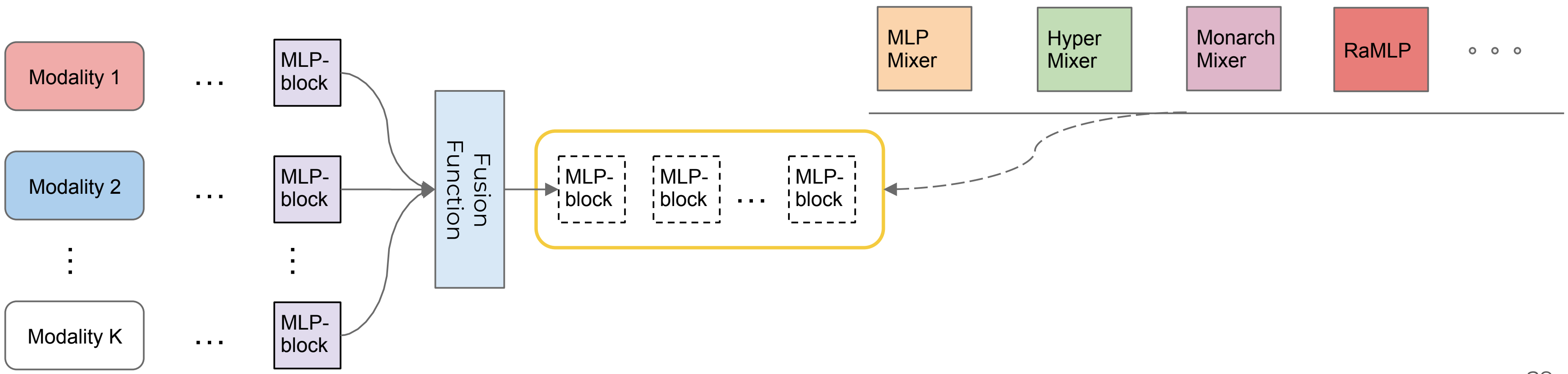
MixMAS: Sampling based micro-benchmarking

1. Take a representative small sample of the dataset [Hogg et al., 2023]
2. Find the best uni-modal encoders for each modality
3. Fix encoders, search for best fusion function



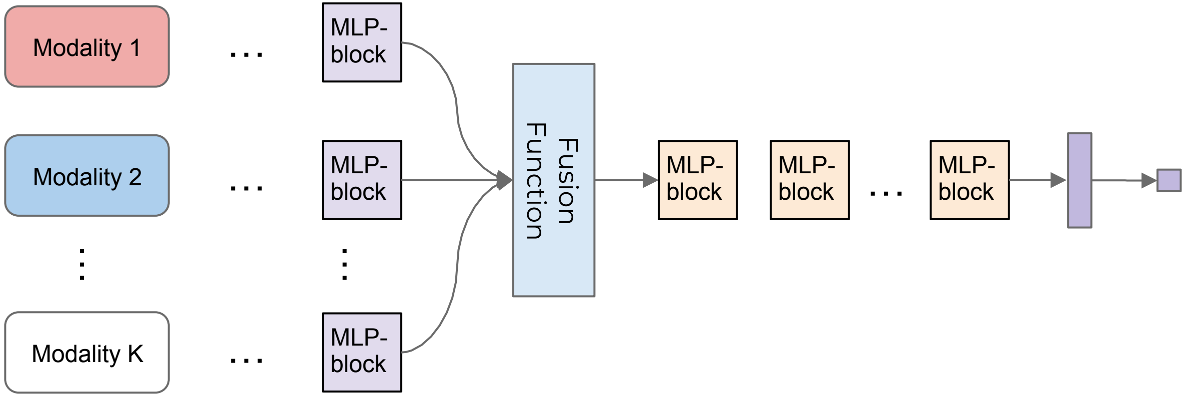
MixMAS: Sampling based micro-benchmarking

1. Take a representative small sample of the dataset [Hogg et al., 2023]
2. Find the best uni-modal encoders for each modality
3. Fix encoders, search for best fusion function
4. Fix Fusion function, search for best fusion network



MixMAS: Sampling based micro-benchmarking

1. Take a representative small sample of the dataset [Hogg et al., 2023]
2. Find the best uni-modal encoders for each modality
3. Fix encoders, search for best fusion function
4. Fix Fusion function, search for best fusion network
5. Train the final model on the whole dataset



Experiments

3 Datasets:	AV-MNIST [Vielzeuf et al., 2018]	MIMIC-III [Johnson et al., 2015]	MM-IMDB [Arevalo, et al., 2017]
Field	Multimedia	Healthcare	Multimedia
Modalities	Image / Audio	Time Series / Tabular	Image / Text
Samples train / val / test	70,000	32,615	36,212

Average of 10 runs

	MM-IMDB		AV-MNIST		MIMIC-III	
Architecture	F1-w. (%) (avg)	Training Params (M)	Acc. (%) (avg)	Training Params (M)	Acc. (%) (avg)	Training Params (M)
M2-Mixer	46.66 ± 0.44	16.7	73.20 ± 0.2	8.3	78.32 ± 0.3	0.029
MixMAS	$\mathbf{49.58 \pm 0.5}$	10.37	$\mathbf{75.79 \pm 0.3}$	9.33	78.3 ± 0.73	0.033

Results of Micro-Benchmarking - Encoder Selection

	MM-IMDB		AV-MNIST		MIMIC-III
Sampling(%)	23%		12%		21%
Module	Score F1-w(%)	Module	Score Acc(%)	Module	Score Acc(%)
Image Encoder Selection		**Image Encoder Selection**		**Time-Series Encoder Selection**	
MLPMixer	**24.02**	MLPMixer	44.27	MLPMixer	40.77
HyperMixer	16.89	**HyperMixer**	**56.15**	**HyperMixer**	**45.36**
RaMLP	14.44	RaMLP	47.52	MonarchMixer	44.38
Text Encoder Selection		**Audio Encoder Selection**		**Tabular Encoder Selection**	
MLPMixer	9.20	MLPMixer	27.40	—	—
HyperMixer	15.07	**HyperMixer**	**29.16**	—	—
MonarchMixer	**28.55**	MonarchMixer	28.49	—	—
Fusion Function Selection		**Fusion Function Selection**		**Fusion Function Selection**	
ConcatFusion	**19.56**	**ConcatFusion**	**18.38**	**ConcatFusion**	**28.55**
MeanFusion	10.20	MeanFusion	9.61	MeanFusion	4.28
MaxFusion	9.07	MaxFusion	6.20	MaxFusion	6.73
Fusion Network Selection		**Fusion Network Selection**		**Fusion Network Selection**	
HyperMixer	**29.0**	**HyperMixer**	**53.47**	**HyperMixer**	**38.15**
MLPMixer	25.97	MLPMixer	42.17	MLPMixer	34.14

Results of Micro-Benchmarking - Fusion Function Selection

MM-IMDB		AV-MNIST		MIMIC-III	
Sampling(%)	23%	12%		21%	
Module	Score F1-w(%)	Module	Score Acc(%)	Module	Score Acc(%)
Image Encoder Selection		**Image Encoder Selection**		**Time-Series Encoder Selection**	
MLPMixer	**24.02**	MLPMixer	44.27	MLPMixer	40.77
HyperMixer	16.89	**HyperMixer**	**56.15**	**HyperMixer**	**45.36**
RaMLP	14.44	RaMLP	47.52	MonarchMixer	44.38
Text Encoder Selection		**Audio Encoder Selection**		**Tabular Encoder Selection**	
MLPMixer	9.20	MLPMixer	27.40	—	—
HyperMixer	15.07	**HyperMixer**	**29.16**	—	—
MonarchMixer	**28.55**	MonarchMixer	28.49	—	—
Fusion Function Selection		**Fusion Function Selection**		**Fusion Function Selection**	
ConcatFusion	**19.56**	**ConcatFusion**	**18.38**	**ConcatFusion**	**28.55**
MeanFusion	10.20	MeanFusion	9.61	MeanFusion	4.28
MaxFusion	9.07	MaxFusion	6.20	MaxFusion	6.73
Fusion Network Selection		**Fusion Network Selection**		**Fusion Network Selection**	
HyperMixer	**29.0**	**HyperMixer**	**53.47**	**HyperMixer**	**38.15**
MLPMixer	25.97	MLPMixer	42.17	MLPMixer	34.14

Results of Micro-Benchmarking - Fusion Network Selection

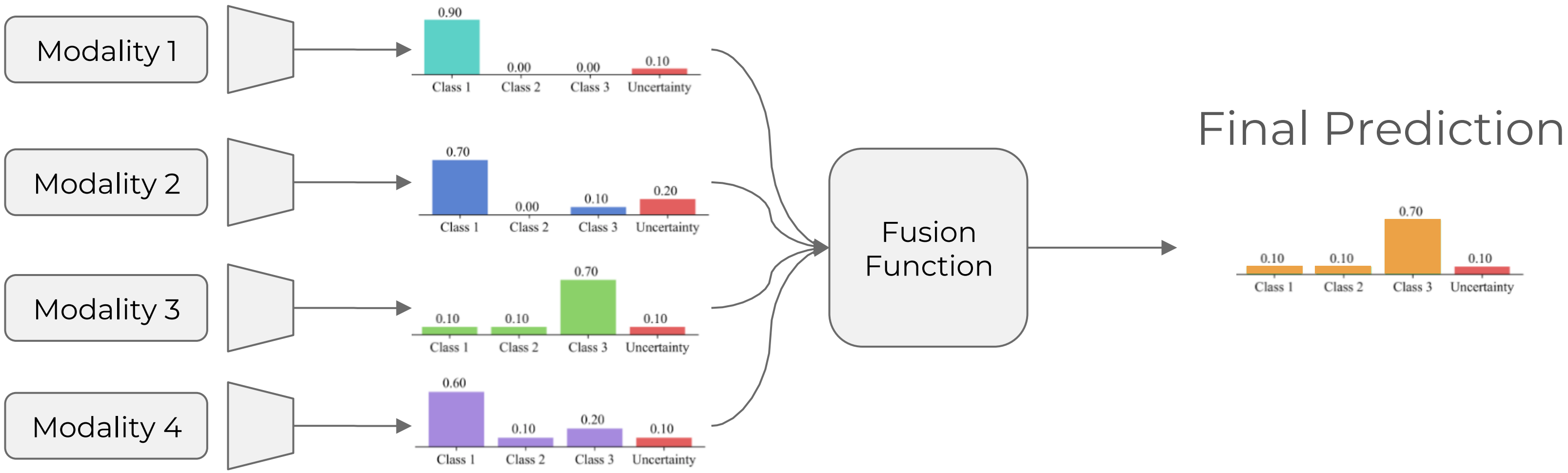
	MM-IMDB		AV-MNIST		MIMIC-III	
Sampling(%)	23%		12%		21%	
	Module	Score F1-w(%)	Module	Score Acc(%)	Module	Score Acc(%)
	Image Encoder Selection		**Image Encoder Selection**		**Time-Series Encoder Selection**	
	MLPMixer	**24.02**	MLPMixer	44.27	MLPMixer	40.77
	HyperMixer	16.89	**HyperMixer**	**56.15**	**HyperMixer**	**45.36**
	RaMLP	14.44	RaMLP	47.52	MonarchMixer	44.38
	Text Encoder Selection		**Audio Encoder Selection**		**Tabular Encoder Selection**	
	MLPMixer	9.20	MLPMixer	27.40	—	—
	HyperMixer	15.07	**HyperMixer**	**29.16**	—	—
	MonarchMixer	**28.55**	MonarchMixer	28.49	—	—
	Fusion Function Selection		**Fusion Function Selection**		**Fusion Function Selection**	
	ConcatFusion	**19.56**	**ConcatFusion**	**18.38**	**ConcatFusion**	**28.55**
	MeanFusion	10.20	MeanFusion	9.61	MeanFusion	4.28
	MaxFusion	9.07	MaxFusion	6.20	MaxFusion	6.73
	Fusion Network Selection		**Fusion Network Selection**		**Fusion Network Selection**	
	HyperMixer	**29.0**	**HyperMixer**	**53.47**	**HyperMixer**	**38.15**
	MLPMixer	25.97	MLPMixer	42.17	MLPMixer	34.14

Our contributions

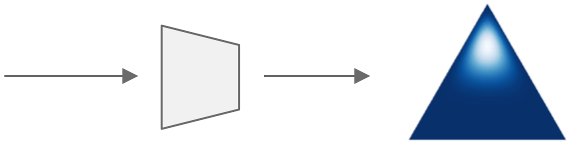
Ph.D. thesis of Grigor Bezirganyan, and collaborators @AMU

1. **M2-Mixer**: Design adaptive, conceptually, computationally simple, scalable multimodal deep learning architecture
2. **MixMax:** Find the optimal multimodal deep learning architecture
3. **DBF: Quantify the uncertainties** in multimodal learning
4. **LUMA:** Provide a benchmark dataset for multimodal learning and uncertainty quantification

Related Work: Multimodal Evidential Deep Learning



Evidential Neural Network



Predict the parameters of Dirichlet Distribution

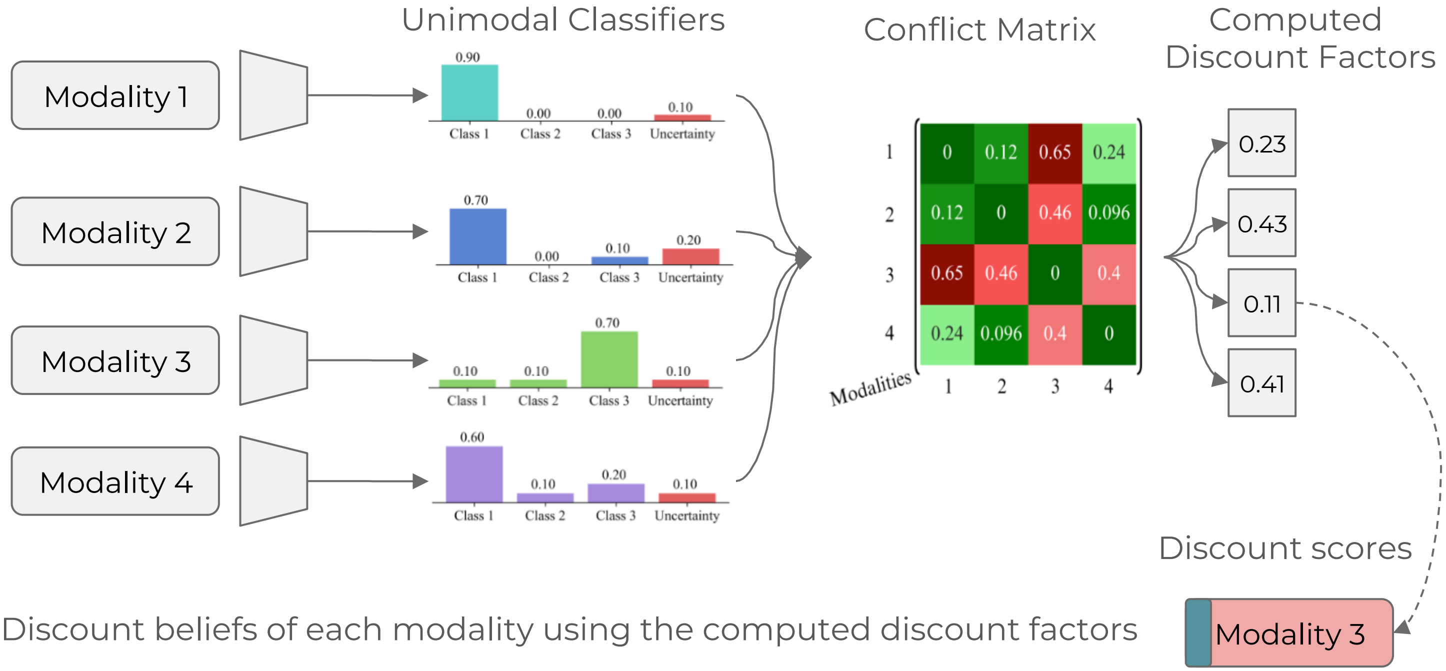
(Han et. al, 2022)

Modalities can often confidently disagree in their decisions

- Decisions made on conflicting data need to be more uncertain
- Modalities that are in conflict with others need to contribute less to the decision

Discount opinions that contradict with lots of other modalities

Discount modalities that contradict with lots of other modalities

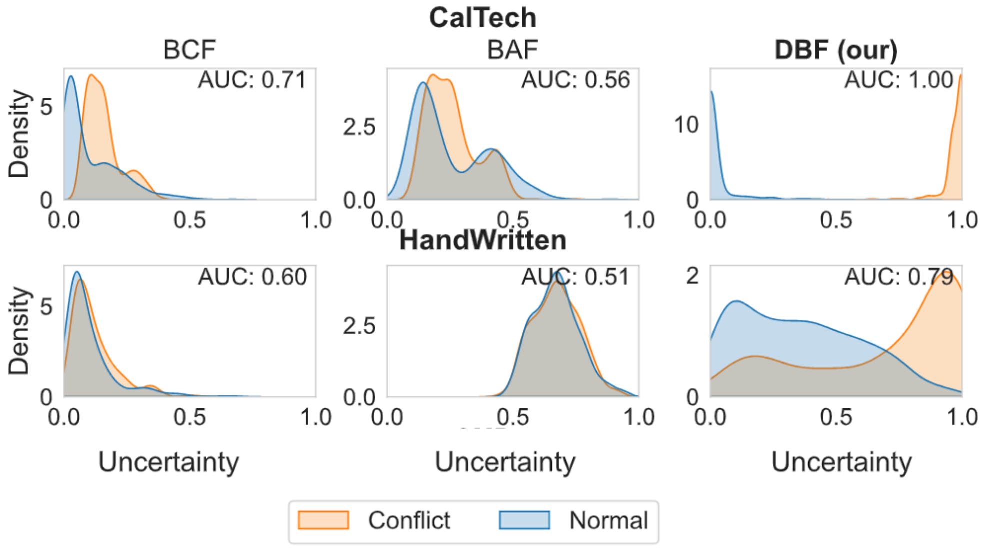


Readjust Uncertainties & Beliefs and Fusion

Discount beliefs of each modality using the computed discount factors

G. Bezirganyan, S. Sellami, L. Berti-ÉQuille and S. Fournier, (2024). Multimodal Learning with Uncertainty Quantification based on Discounted Belief Fusion. arXiv preprint arXiv:2412.18024. (Accepted to AIStats 2025)

Discounting Belief Fusion effectively distinguishes between conflictive and non-conflictive modalities



G. Bezirganyan, S. Sellami, L. Berti-ÉQuille and S. Fournier, (2024). Multimodal Learning with Uncertainty Quantification based on Discounted Belief Fusion. arXiv preprint arXiv:2412.18024. (Accepted to AIStats 2025)

Our contributions

Ph.D. thesis of Grigor Bezirganyan, and collaborators @AMU

1. **M2-Mixer**: Design adaptive, conceptually, computationally simple, scalable multimodal deep learning architecture
2. **MixMax:** Find the optimal multimodal deep learning architecture
3. **DBF:** Quantify the uncertainties in multimodal learning
4. **LUMA:** Provide a benchmark dataset for multimodal learning and uncertainty quantification

Existing Multimodal Datasets

- Lack the ability to inject controlled amount of noise in each modality
- Injected noises are artificial and do not reflect real-life scenarios
- Not enough samples in the datasets

LUMA: Benchmark Dataset for Learning from Uncertain and Multimodal Data

Image

Text

Audio

24000 Images collected from CIFAR-100 Dataset

~50000 Texts Generated with Large Language Models

~130000 Audio samples extracted from various speech corpuses

Grigor Bezirganyan, Sana Sellami, Laure Berti-Equille, and Sebastien Fournier. Luma: A benchmark dataset for learning from uncertain and multimodal data. arXiv preprint arXiv:2406.09864,2024

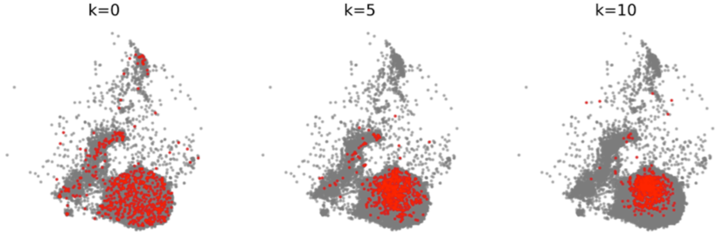
LUMA: Benchmark Dataset for Learning from Uncertain and Multimodal Data

Image	Text	Audio
	I was riding my beautiful black stallion named Shadow, through the park yesterday. It was a sunny day, and the wind was blowing in my hair. I felt free and happy.	Pronunciation of word "Horse"

Grigor Bezirganyan, Sana Sellami, Laure Berti-Equille, and Sebastien Fournier. Luma: A benchmark dataset for learning from uncertain and multimodal data. arXiv preprint arXiv:2406.09864,2024

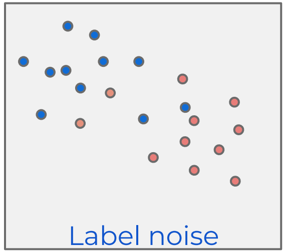
Adding Noises to LUMA:

1. Diversity

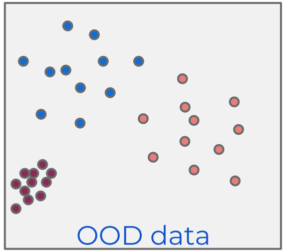


2. Sample Noise

3. Label Noise



4. OOD Injection



Grigor Bezirganyan, Sana Sellami, Laure Berti-Equille, and Sebastien Fournier. Luma: A benchmark dataset for learning from uncertain and multimodal data. arXiv preprint arXiv:2406.09864,2024

LUMA: Benchmark Dataset for Learning from Uncertain and Multimodal Data

Method	Clean		↘ Diversity		↗ Label Noise		↗ Sample Noise	
	Ale.	Epi.	Ale.	Epi.	Ale.	Epi.	Ale.	Epi.
MCD Image	1.00	1.03	-15.73%	-11.66%	+59.20%	+54.51%	+4.44%	+2.18%
MCD Audio	0.52	0.70	-5.54%	+2.16%	+96.63%	+54.49%	+23.12%	+14.40%
MCD Text	0.37	1.01	-3.91%	-2.62%	+93.59%	+2.41%	+64.96%	-2.03%
MCD Multi.	0.26	0.78	-8.52%	-1.21%	+122.44%	+11.60%	+59.14%	+9.89%
DE Image	1.45	1.40	-37.49%	-8.54%	-7.43%	+0.24%	-18.46%	-3.22%
DE Audio	0.56	0.99	-27.39%	-3.34%	**+156.40%**	+50.43%	**+70.26%**	+34.41%
DE Text	0.42	1.01	+5.02%	-6.15%	+81.26%	-0.51%	+62.24%	-7.11%
DE Multi.	0.31	0.82	-22.80%	-3.40%	+115.15%	+20.62%	+45.97%	+5.54%
RCML Multi.	1.99	0.43	**+8.34%**	**+16.16%**	+64.72%	**+106.16%**	+36.19%	**+58.21%**

Grigor Bezirganyan, Sana Sellami, Laure Berti-Equille, and Sebastien Fournier. Luma: A benchmark dataset for learning from uncertain and multimodal data. arXiv preprint arXiv:2406.09864,2024

Conclusion & Future Work (1/2)

- **M2-Mixer**: an all-MLP based architecture for multimodal fusion with multi-head loss to Improve modality representations
- **MixMAS**: a sampling based micro-benchmarking pipeline for mixer based architecture search
- **DBF:** a Discount Belief Approach for uncertainty quantification in multimodal classification
- **LUMA:** a benchmarking dataset for uncertainty quantification in multimodal settings

- Model hybridation / ensembling architectures
- Add more advanced sampling strategies (e.g. uncertainty based sampling)

Conclusion & Future Work (2/2)

Learning from multimodal data offer **new challenges for data and model engineering R&D**

Requires **interdisciplinary research**:
- DB, ML, Statistics
- Modality-dependent expertise (e.g., remote sensing, audio signal processing, computer vision, etc.)
- Application-dependent expertise (climate, biology, healthcare, etc.)

Requires **humans in the loop** orchestration with higher degree of complexity

There are many **research opportunities** for:
- Managing and orchestration human/machine or agent resources
- Revisiting our methods & technologies to leverage multimodal data

Thanks!

References

1. Wang, C., Liu, X., Yue, Y., Tang, X., Zhang, T., Jiayang, C., ... & Zhang, Y. (2023). Survey on factuality in large language models: Knowledge, retrieval and domain-specificity. arXiv preprint arXiv:2310.07521
2. Munn, L., Magee, L., & Arora, V. (2023). Truth Machines: Synthesizing Veracity in AI Language Models. arXiv preprint arXiv:2301.12066.
3. Fadeeva, E., Vashurin, R., Tsvigun, A., Vazhentsev, A., Petrakov, S., Fedyanin, K., ... & Shelmanov, A. (2023). LM-Polygraph: Uncertainty Estimation for Language Models. arXiv preprint arXiv:2311.07383.
4. James Thorne, Andreas Vlachos, Christos Christodoulopoulos, and Arpit Mittal. 2019. Evaluating adversarial attacks against multiple fact verification systems. In Proc. of the 2019 EMNLP-IJCNLP, pages 2944–2953, Hong Kong, China. https://aclanthology.org/D19-1292
5. Atanasova, P., Wright, D., & Augenstein, I. (2020). Generating label cohesive and well-formed adversarial claims. arXiv preprint arXiv:2009.08205. https://github.com/copenlu/fever-adversarial-attacks
6. Gao, J., Hoffmann, H. F., Oikonomou, S., Kiskovski, D., & Bandhakavi, A. (2021). Logically at Factify 2022: Multimodal Fact Verification. arXiv preprint arXiv:2112.09253. https://arxiv.org/abs/2112.09253
7. Verschuuren, P. J., Gao, J., van Eeden, A., Oikonomou, S., & Bandhakavi, A. (2023). Logically at Factify 2023: A Multi-Modal Fact Checking System Based on Evidence Retrieval techniques and Transformer Encoder Architecture. arXiv preprint arXiv:2301.03127. https://arxiv.org/abs/2301.03127
8. FEVER 2.0 Adversarial Attacks Dataset, https://fever.ai/dataset/adversarial.html
9. Defactify Workshop https://aiisc.ai/defactify/
10. Factify Multi-Modal Fact Verification dataset, https://competitions.codalab.org/competitions/35153
11. Dan Saattrup Nielsen and Ryan McConville. “MuMiN: A Large-Scale Multilingual Multimodal Fact-Checked Misinformation Social Network Dataset.” arXiv preprint arXiv:2202.11684 (2022). https://mumin-dataset.github.io/
12. L Berti-Equille, ML Ba. Veracity of big data: challenges of cross-modal truth discovery. Journal of Data and Information Quality (JDIQ) 7 (3), 1-3

References

[Toliskin et al., 2021] I. O. Tolstikhin, N. Houlsby, A. Kolesnikov, L. Beyer, X. Zhai, T. Unterthiner, J. Yung, A. Steiner, D. Keysers, J. Uszkoreit, M. Lucic, and A. Dosovitskiy. Mlp-mixer: An all-mlp architecture for vision. In NeurIPS 34, pages 24261–24272, 2021.

[Vielzeuf et al., 2018] V. Vielzeuf, A. Lechervy, S. Pateux, and F. Jurie. Centralnet: A multilayer approach for multimodal fusion. In Computer Vision - ECCV 2018 Workshops, Springer, 2018.

[Johnson et al., 2015] A. Johnson, T. Pollard, and R. Mark. MIMIC-III Clinical Database, 2015.

[Liang et al., 2021] P. P. Liang, Y. Lyu, X. Fan, Z. Wu, Y. Cheng, J. Wu, L. Chen, P. Wu, M. A. Lee, Y. Zhu, R. Salakhutdinov, and L. Morency. Multibench: Multiscale benchmarks for multimodal representation learning. In Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks 1, 2021.

[Liu et al., 2018] Z. Liu, Y. Shen, V. B. Lakshminarasimhan, P. P. Liang, A. Bagher Zadeh, and L.-P. Morency. Efficient Low-rank Multimodal Fusion With Modality-Specific Factors. In Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 2247–2256. 2018.

[Pérez-Rúa et al., 2019] J. Pérez-Rúa, V. Vielzeuf, S. Pateux, M. Baccouche, and F. Jurie. MFAS: multimodal fusion architecture search. In IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019, Long Beach, CA, USA, June 16-20, 2019, pages 6966–6975. Computer Vision Foundation / IEEE, 2019.

[Sankaran et al., 2021] S. Sankaran, D. Yang, and S.-N. Lim. Multimodal Fusion Refiner Networks, Apr. 2021. arXiv:2104.03435 [cs], unpublished.

[Wu et al., 2018] M. Wu and N. D. Goodman. Multimodal generative models for scalable weakly-supervised learning. In Advances in Neural Information Processing Systems 31, pages 5580–5590, 2018.

[Tsai et al., 2019] Y. H. Tsai, P. P. Liang, A. Zadeh, L. Morency, and R. Salakhutdinov. Learning factorized multimodal representations. In 7th International Conference on Learning Representations, ICLR, 2019.

[Sun et al., 2020] Z. Sun, P. K. Sarma, W. A. Sethares, and Y. Liang. Learning relationships between text, audio, and video via deep canonical correlation for multimodal language analysis. In The 34 AAAI Conference on Artificial Intelligence, 2020.

[Jayakumar et al., 2020] S. M. Jayakumar, W. M. Czarnecki, J. Menick, J. Schwarz, J. W. Rae, S. Osindero, Y. W. Teh, T. Harley, and R. Pascanu. Multiplicative interactions and where to find them. In 8th International Conference on Learning Representations, ICLR 2020, 2020.

[Wang et al., 2020] W. Wang, D. Tran, and M. Feiszli. What makes training multi-modal classification networks hard? In 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR, 2020