

AI/ML AWG Research Request for Collaboration

James Casaletto
12.September.2023

Motivation

- **RR9**

- Newly released $\Rightarrow$ opportunity to publish
- Newly released $\Rightarrow$ how easy it is it to use?
- Has omic + phenotype data $\Rightarrow$ opportunity to do supervised ML

- **OSDR**

- Would benefit from AI/ML tooling/pipelines
- Lots of data are unlabelled
- Lots of data are $p \gg n$

- **Eyesight**

- Mission-critical
- SANS = space-associated neuro-ocular syndrome
- Microgravity $\Rightarrow$ fluid pressure change not the only cause

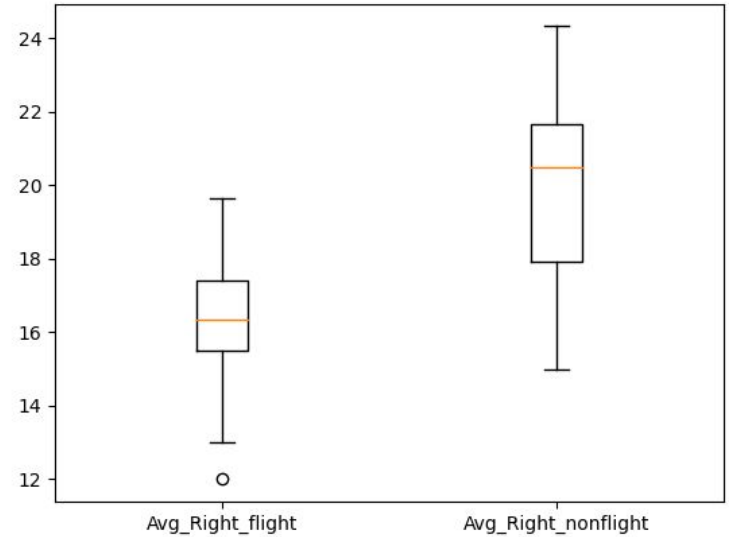
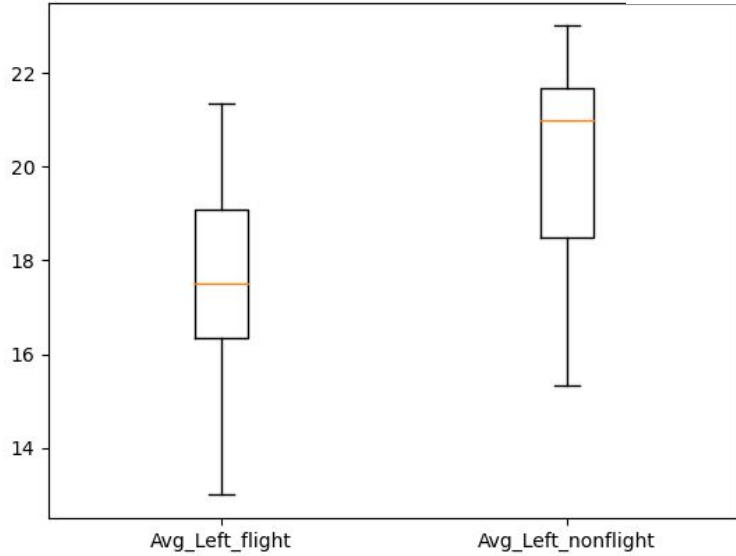
- **Invariant classification**

- Invariance used as proxy for robust correlation and causality
- Invariant classification is good classification
- Not difficult to incorporate into algorithms

Data (RR9)

OSD#	Material Type	Omics data	Pheno-Physio
<u>255</u>	Retina	RNAseq (8 GC + 8 FLT)	
<u>568</u>	Retina		Immunostaining Microscopy
<u>557</u>	Eye/Retina		mCT, histology, immunostaining
<u>583</u>	Eye		Intraocular Pressure
<u>478</u>	Behavior		Gait
397	Retina	RNAseq, methylation (17 simulated mg/rad)	
203	Retina	RNAseq, methylation (60 simulated mg/rad)	
194	Retina	RNAseq (4 BSL, 4 GC, 5 FLT)	
87	Retina	RNAseq (6 FLT)	

Intra-ocular pressure



Tasks

1. Evaluate OSDR data for usability in AI/ML project

- Data types (omics, phenotypic, etc)
- Naming conventions (sample names, file names, ... etc)
- Write a report / can be used beyond this project

2. Design/implement/leverage a data labeler

- 96 unlabeled retina RNA-seq samples (16 labeled)
- Required for supervised ML algorithms (e.g. regression, classification)
- Can be used beyond this project

3. Design/implement/leverage covariate-aware data generator

- ML algorithms will need lots more data than what we have (by orders of magnitude)
- Have prototype working
- Can be used beyond this project

4. Implement invariance in regression and classification

- Invariance combats spurious correlations
- Algorithm change involves how training, testing, and validation data are selected
- Have prototype working / can be used beyond this project

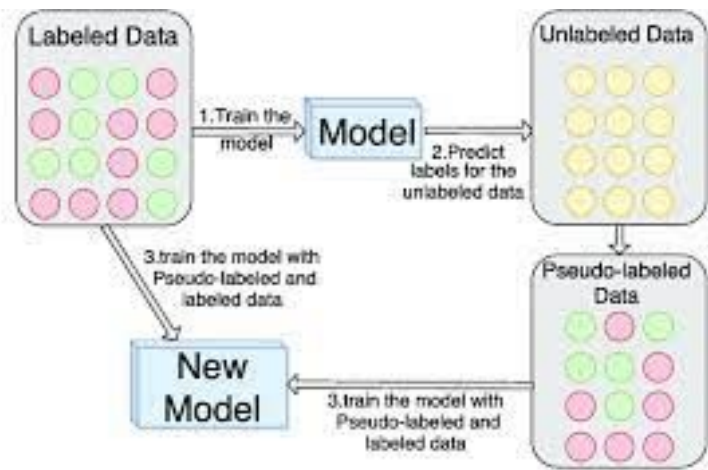
5. Implement digital twin

- Stretch goal to support what-if analysis (manipulate features $\Rightarrow$ examine phenotype)

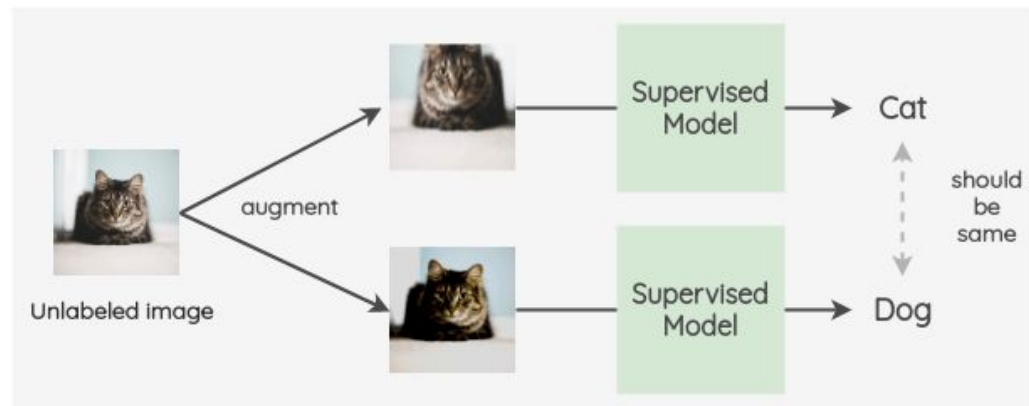
Some more information on some of the tasks

Task: 2. Design/implement/leverage a data labeler

pseudo-labeling



consistency regularization



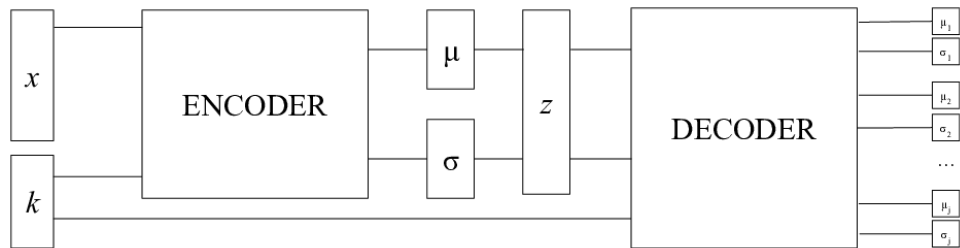
Other approaches:

- Few-shot learning
- Commercial tools like Label Studio

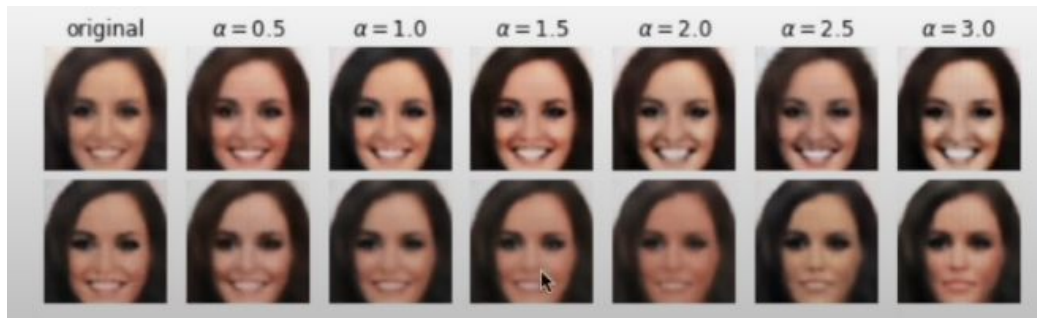
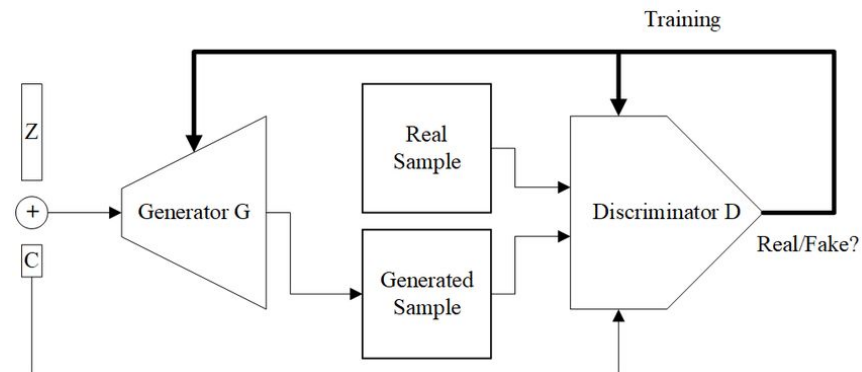
Task 3: Design/implement/leverage covariate-aware data generator

Task 5: Implement digital twin

conditional variational autoencoder (CVAE)



conditional generative adversarial network (CGAN)



Task 4: Implement invariance in regression and classification

