# Data Centric AI for QEC (AI4Q)

Feature Engineering for Traditional ML model decoding of QEC on real data.

Aadi Patwardhan, Supervisors : Rihan Hai, Tim Littau

## 1. Motivation

Why I did this?

Quantum Computing struggles to scale due to lack of fault tolerant hardware : Needs Quantum Error Correction (QEC)!

Existing works use deep learning and simulated data.

QEC data model for simple traditional ML QEC decoder?

To what extent can feature engineering and temporal aggregation enable traditional ML models to decode QEC on real hardware data?

Empirical Study of Feature Engineering:

**Need, Competence, Priority.**

Q1 : What feature groups aid traditional ML models to solve the QEC problem?
Q2 : How do temporal aggregation-based data models compare with existing baselines?
Q3 : Which features do traditional ML models learn to use?

Groundwork in larger Model Lake Vision with data lake system for QEC data and downstream zoo of ML models for the various QEC tasks and setups.

## 2. Methodology

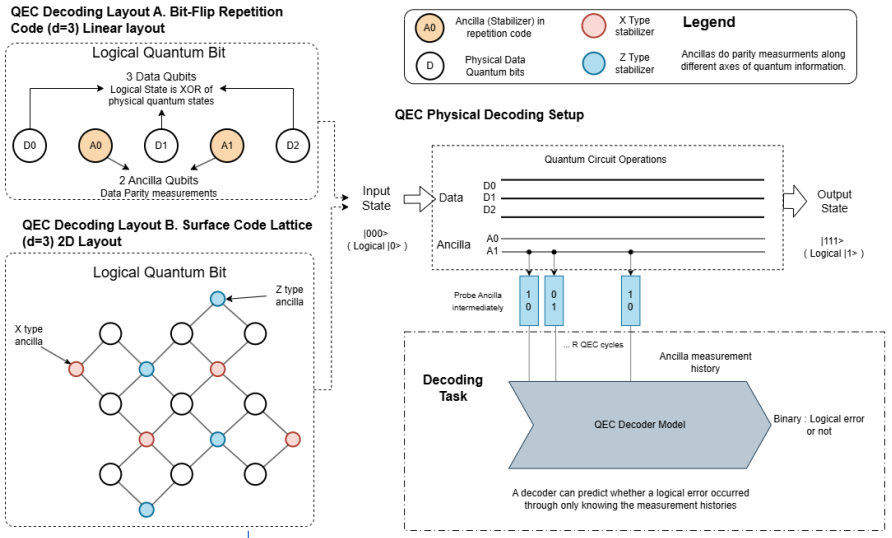
What I did?

Built a data processing pipeline to enable Traditional Models trained to do QEC decoding.

Evaluated against classical MWPM and deep LSTM Baseline from QuTech on LEP, LER and test set Accuracy.

**Empirical Analysis :** Ablations (Need), Accuracy decay (Competence) and Feature importance(Priority/usage).

## QEC Zoom in



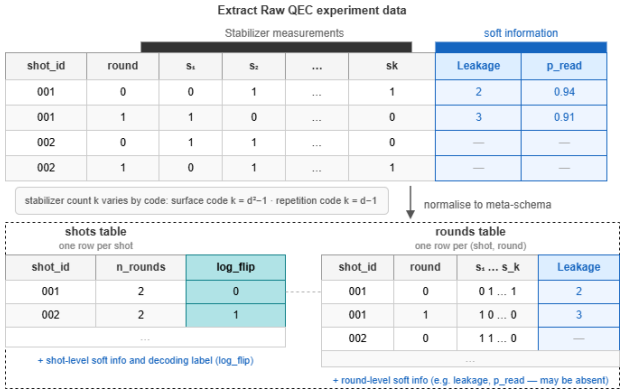

## QEC Lake Pipeline

Data Lake for AI enhanced QEC data processing (**VLDB**)

**Processed 3 REAL datasets:**

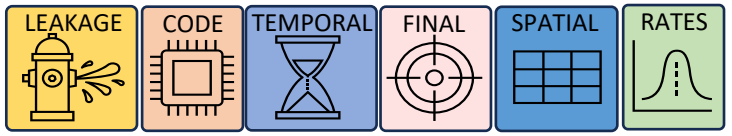
Repetition Code Regime :
- DiCarlo Lab QuTech

Surface Code Regime :
- Google Sycamore d3
- Google Sycamore d5.




## Feature Engineering :

Temporal Aggregation of measurement history (QEC Data).




## Mini Model Zoo

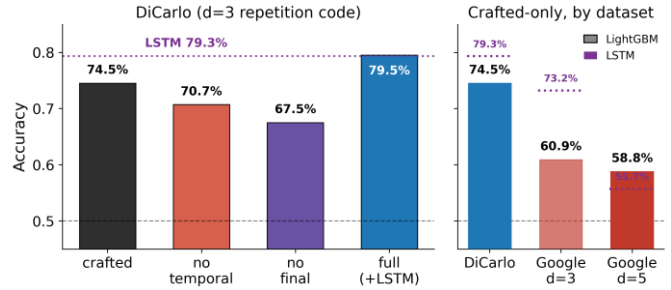
Gradient Boosting Tree Models: CatBoost, LightGBM, XGBoost
Linear : Logistic Regression,
Deep : LSTM
Pipeline also supports (SLOW): Torch MLP, Random Forest and SVM.

## 3. Results & Take-aways

What I learned?

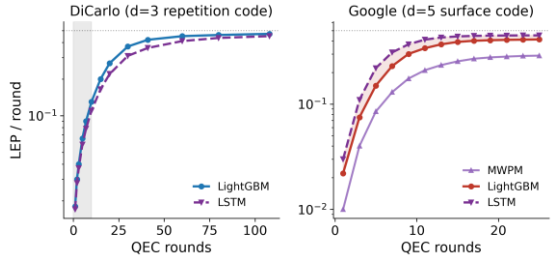
### Ablation Study

Take-away : Final Information is most important. Temporal is direction to improve with LSTM encodings. Complexity of code and dataset size are critical.



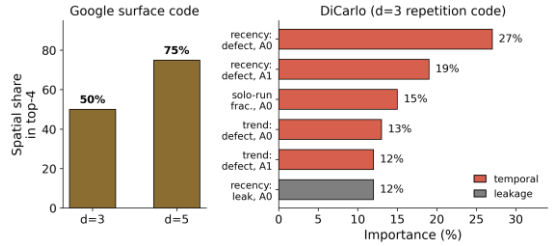

### Accuracy Decay

Take-away : Aggregation competitive in initial (~10) rounds.




### Feature Importance

Take-away : Use spatial class for larger surface codes. Recency and final features are prioritised.




Traditional ML is competitive on smaller datasets and simpler QEC codes at fewer cycles. Mini Model lake displays how even simple interpretable models lay the groundwork for data driven AI based QEC decoding.