Curation before Creation

Guided Sampling for Greener HPC

Dr Amanda J Parker

School of Computing

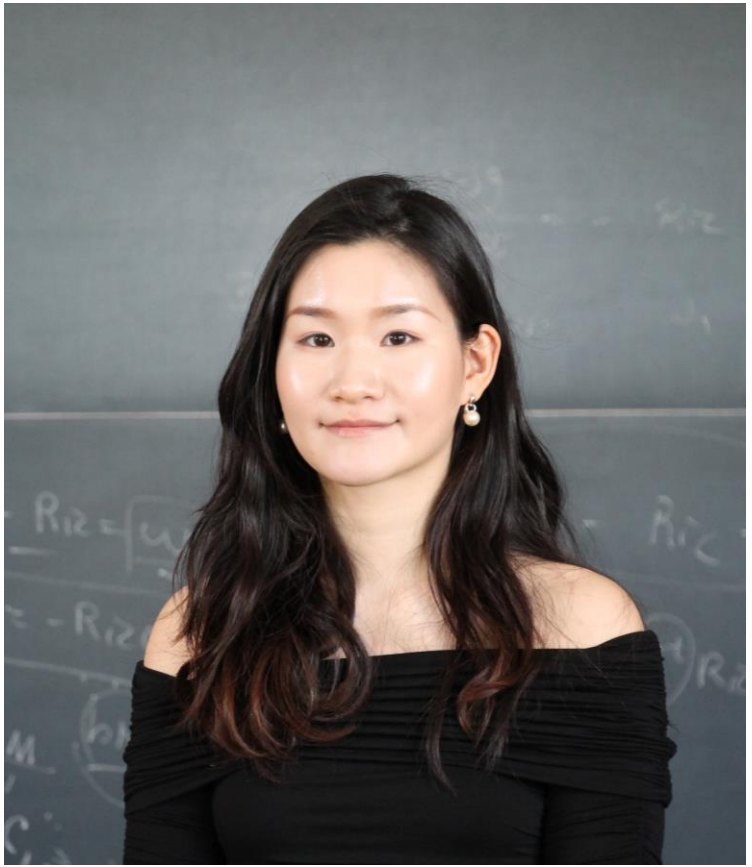
Australian National University

Australian National University

Curation before Creation: Guided Sampling for Greener HPC

Supported by:
ANU Integrated AI Network
Computing for Social Good Seed Grant

Researchers:
Ms. Chloe Lin
Ms. Haiqi Dong
Dr. Amanda J. Parker



Ms. Chloe Lin

Research Context

Machine Learning Mismatches

- Big Data Methods vs Small Data Applications
- Optimised Model Training Speed vs Model Performance
- Abundant Unlabelled Data vs Expensive Labelled Data
- Complexity vs Memory (e.g. randomness, asymmetry)
- System Size Variation vs Tabular ML Methods
- Domain knowledge vs Bias

Research Context

Machine Learning Mismatches

- Big Data Methods vs Small Data Applications
- Optimised Model Training Speed vs Model Performance
- Abundant Unlabelled Data vs Expensive Labelled Data
- Complexity vs Memory (e.g. randomness, asymmetry)
- System Size Variation vs Tabular ML Methods
- Domain knowledge vs Bias

ML methods developed with these settings in mind

Computational Science

Costs

- Direct Resource costs
- Opportunity costs
- Environmental costs

Benefits

- New Technologies
- Health
- Environmental Goods

Finding Balance
Merit Allocation Schemes

- Scientific Merit
- Benefit and Impact
- Investigator
- Feasibility

Finding Balance:
Merit Allocation Schemes

- Scientific Merit
- Benefit and Impact
- Investigator
- Feasibility

Where is the Gap?

Finding Balance
Green HPC

- Hardware Efficiencies
- Software Efficiencies

Finding Balance
Green HPC

Assume Fixed

- Hardware Efficiencies
- Software Efficiencies

Finding Balance
Guided Sampling

Assume Fixed

- Hardware Efficiencies
- Software Efficiencies

Do less Experiments!

Finding Balance
Guided Sampling

Assume Fixed

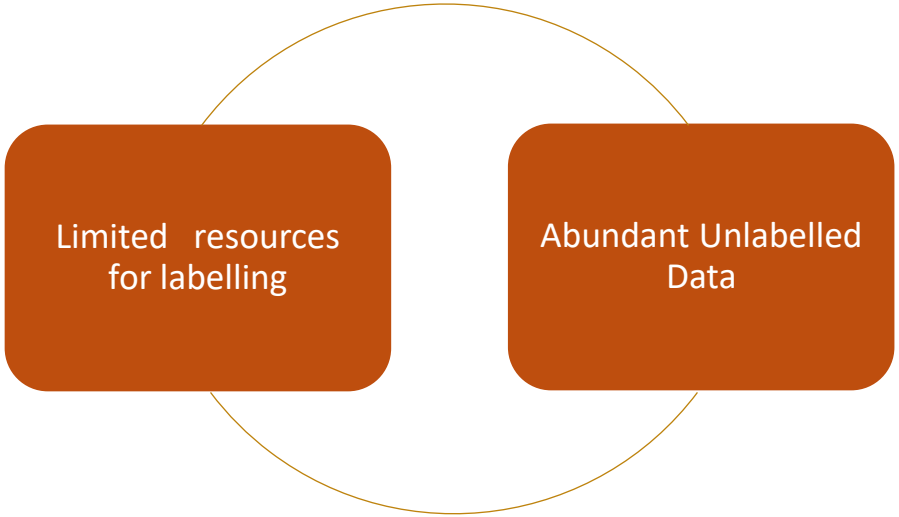
- Hardware Efficiencies
- Software Efficiencies

~~Do less Experiments!~~

Optimise which experiments you conduct relative to their cost

Addressing ML Mismatches: Scientific Data

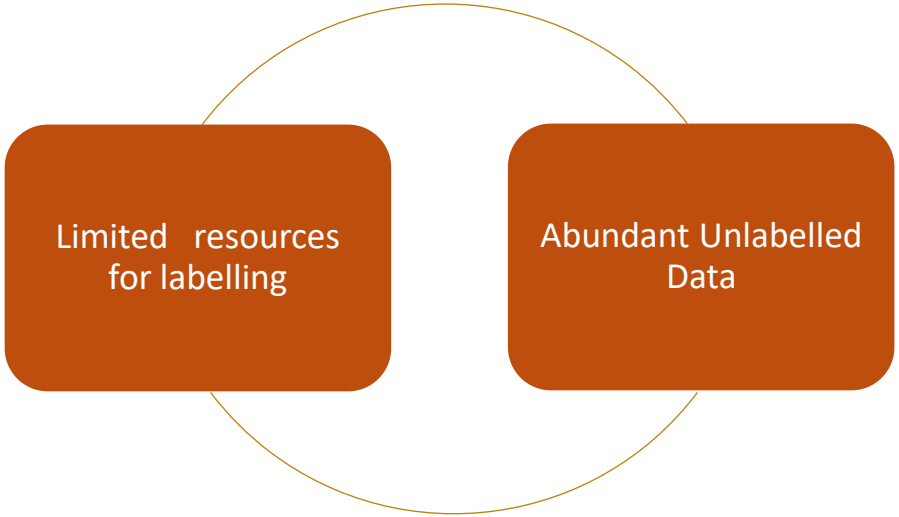
Acquisition of scientific data can be expensive and time-consuming.
We often face:



In a scientific context could require an experiment, simulation or measurement

Addressing ML Mismatches: Scientific Data

Acquisition of scientific data can be expensive and time-consuming.
We often face:

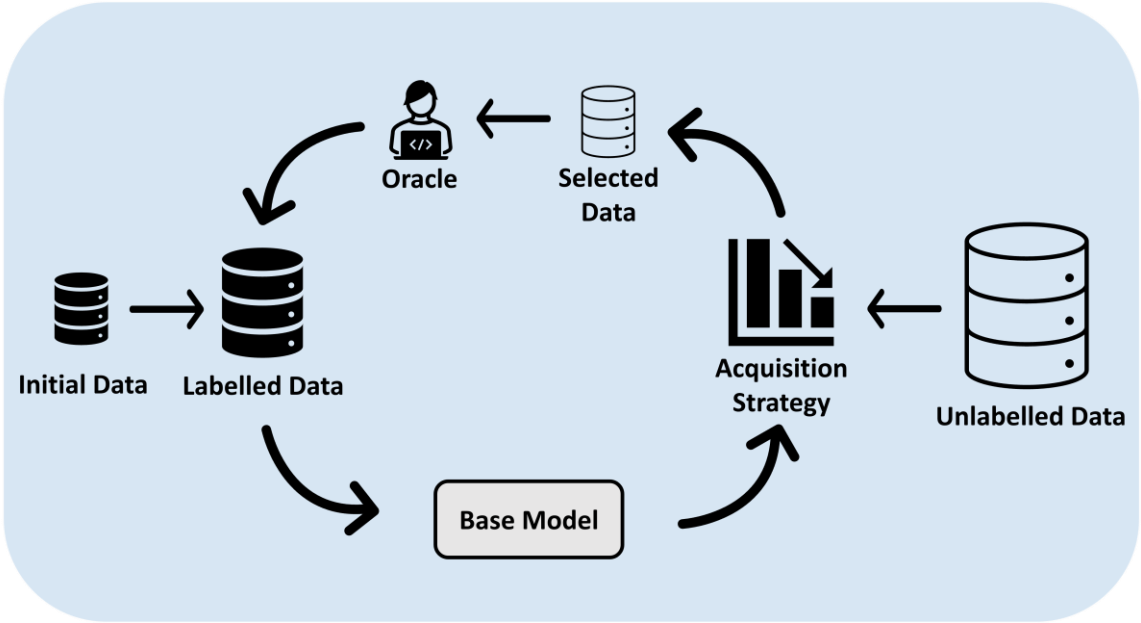


Unsupervised Learning
Guided Sampling

In a scientific context could require an experiment, simulation or measurement

Sampling Strategies

Pool Active Learning

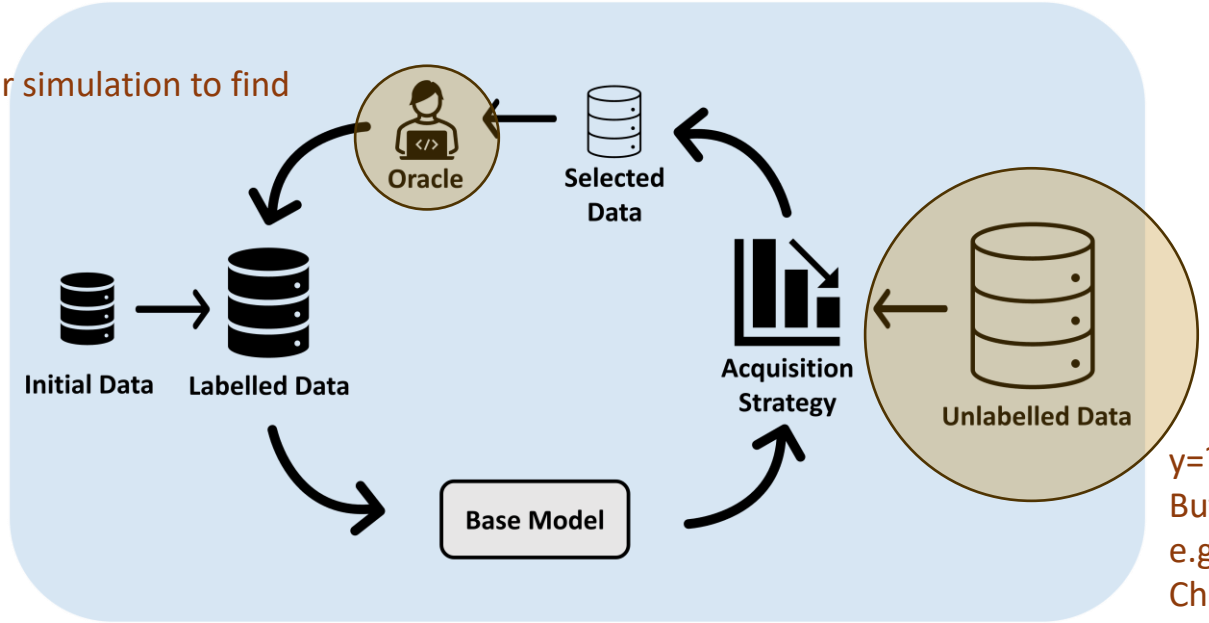


Annotated data with known labels is shown as filled (black) bins, and unlabelled feature data is shown as unfilled bins. AL is initialised with a large pool of unlabelled data (right) and a small set of labelled data (left). Oracles represent experiments, measurements, or simulations performed to obtain numerical or categorical labels when annotating data. The acquisition strategy (selector) ranks all unlabelled data. The base model is the selected predictive model. A single unlabelled data sample or batch of samples is selected for annotation in each AL iteration.

Sampling Strategies

Pool Active Learning

Experiment or simulation to find label y



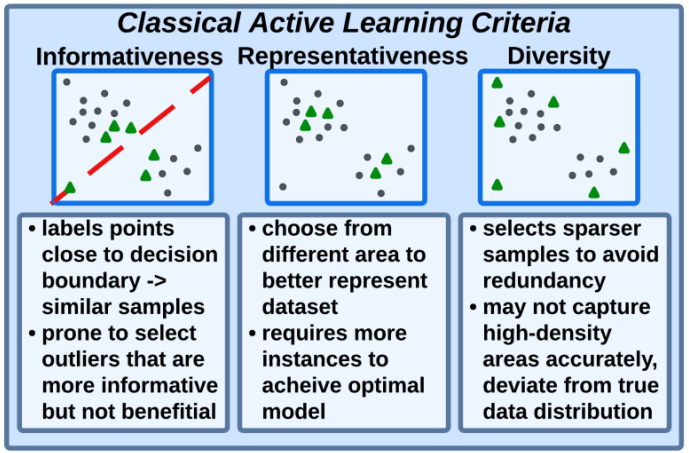
y=?
But fully characterise
e.g. xyz co-ordinates
Chemical constituent

Annotated data with known labels is shown as filled (black) bins, and unlabelled feature data is shown as unfilled bins. AL is initialised with a large pool of unlabelled data (right) and a small set of labelled data (left). Oracles represent experiments, measurements, or simulations performed to obtain numerical or categorical labels when annotating data. The acquisition strategy (selector) ranks all unlabelled data. The base model is the selected predictive model. A single unlabelled data sample or batch of samples is selected for annotation in each AL iteration.

Sampling Strategies

Unlabelled → **Brute Force** → **Labelled**

Random
Quasi-Random
Greedy
QBC



Regression Driven Approach

- **Practical Applications in Scientific Contexts**
  - Scientific experiments, measurements, and simulations involve higher labelling costs and yield numerical outputs
- **Limited Research & Development**
  - Development of Active Learning strategies for regression lags behind those designed for classification
- **Green Computing**
  - Reduces training set size and minimises number of experiments needed in scientific research

Sampling Strategies for Scientific Data

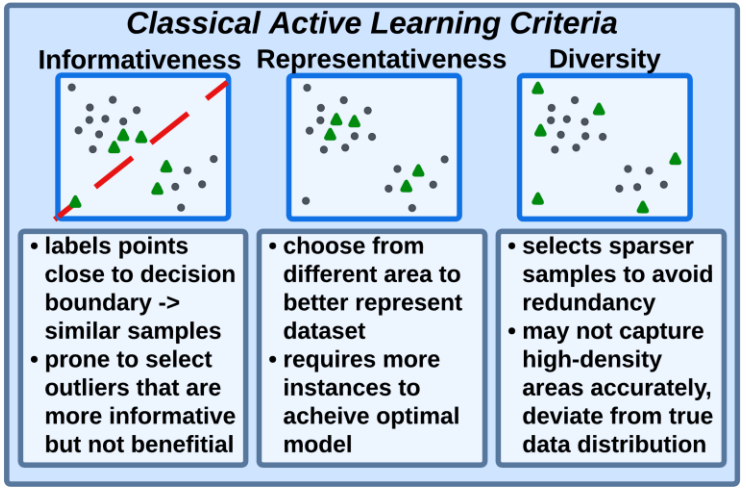
Regression Driven Approach

Practical Applications in Scientific Contexts
- Scientific experiments, measurements, and simulations involve higher labelling costs and yield numerical outputs

Limited Research & Development
- Development of Active Learning strategies for regression lags behind those designed for classification

Green Computing
- Reduces training set size and minimises number of experiments needed in scientific research

Active learning query strategies typically prioritise: informativeness, representativeness or diversity of sampled data



Other Scientific Data Challenges

- Model driven or Data driven
  - Unbalanced Data
  - Small Data sets
- Small initially labelled set

IAI Seed Grant

- *A fixed sample of publicly available materials science data sets*
  - *NCI ANUMAS projects 2019-2021*
  - *61 Projects*
  - *93 Papers*
  - *39 accessible data sets (8 CIs) NOT necessarily collected to train a predictive model*

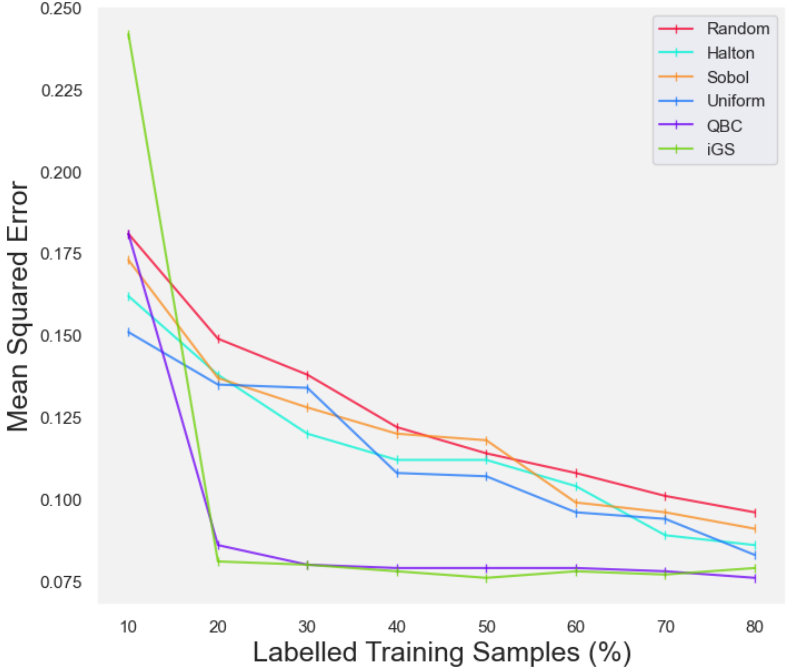
- *Subsample appropriate data sets with variety of cross domain and machine learning informed sampling methods.*

- *Increased benefits (model improvement) benchmarked against increased costs (more calculations)*

  Learning Curves

Quantifying Cost Vs Model Improvement

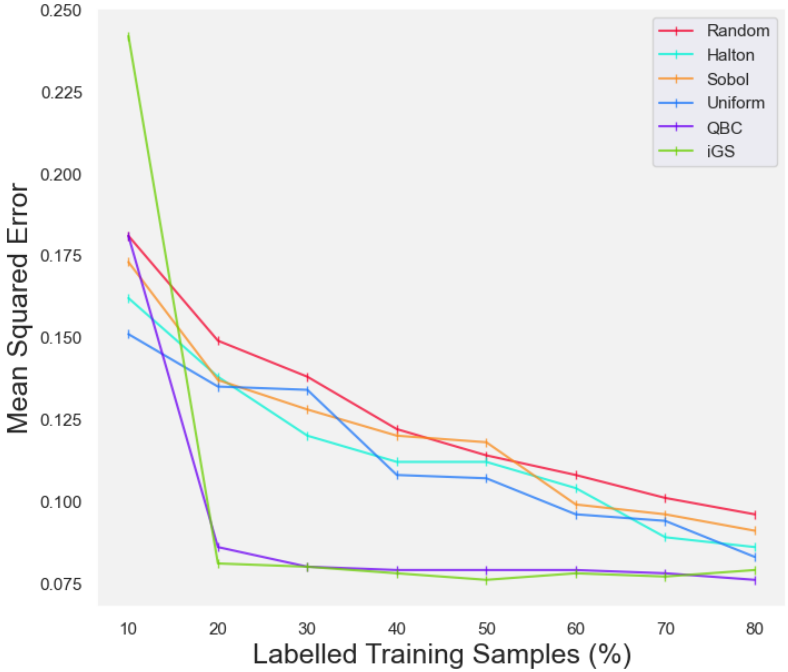
Hold out test set learning curves



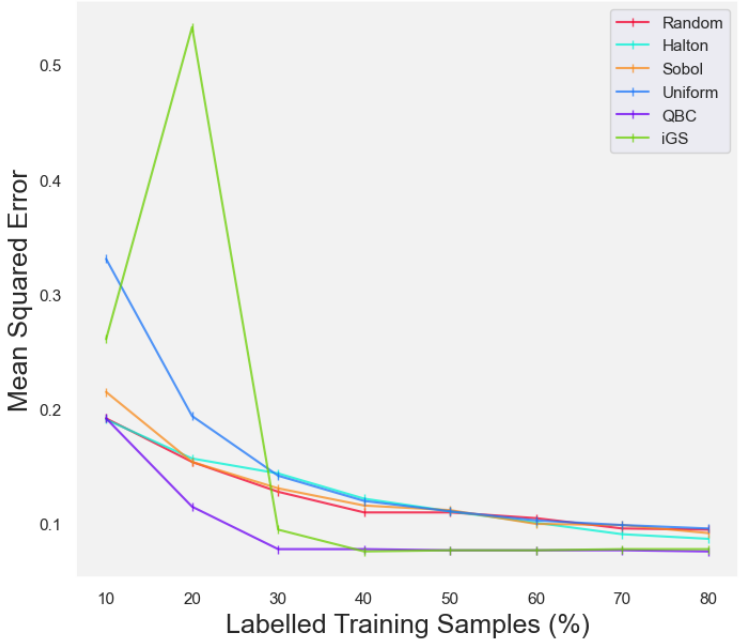
Palladium Nanoparticles Defect Energy Opetal et al.

Quantifying Cost Vs Model Improvement

Hold out test set learning curves



Palladium Nanoparticles Defect Energy Opetal et al.



Silver Nanoparticles Band Gap Barnard et al.

Summary

- Identify ML goals at project conception
  - Do you want to fit a predictive model?
  - YES – guided sampling

- Extract as much insight from available data to direct future experiments
  - Metrics to justify experimental costs!
  - Correlations and unsupervised ML are powerful
  - Avoid biases, new science

- Future Research
  - Benchmarking and a pipeline for selecting methods (Is your data a good fit for guided sampling?)