How to Run ML Experiments

Beyond CS 181 - Lecture 9

Anna Trella

HARVARD
John A. Paulson
School of Engineering
and Applied Sciences

Welcome Back!

General Overview

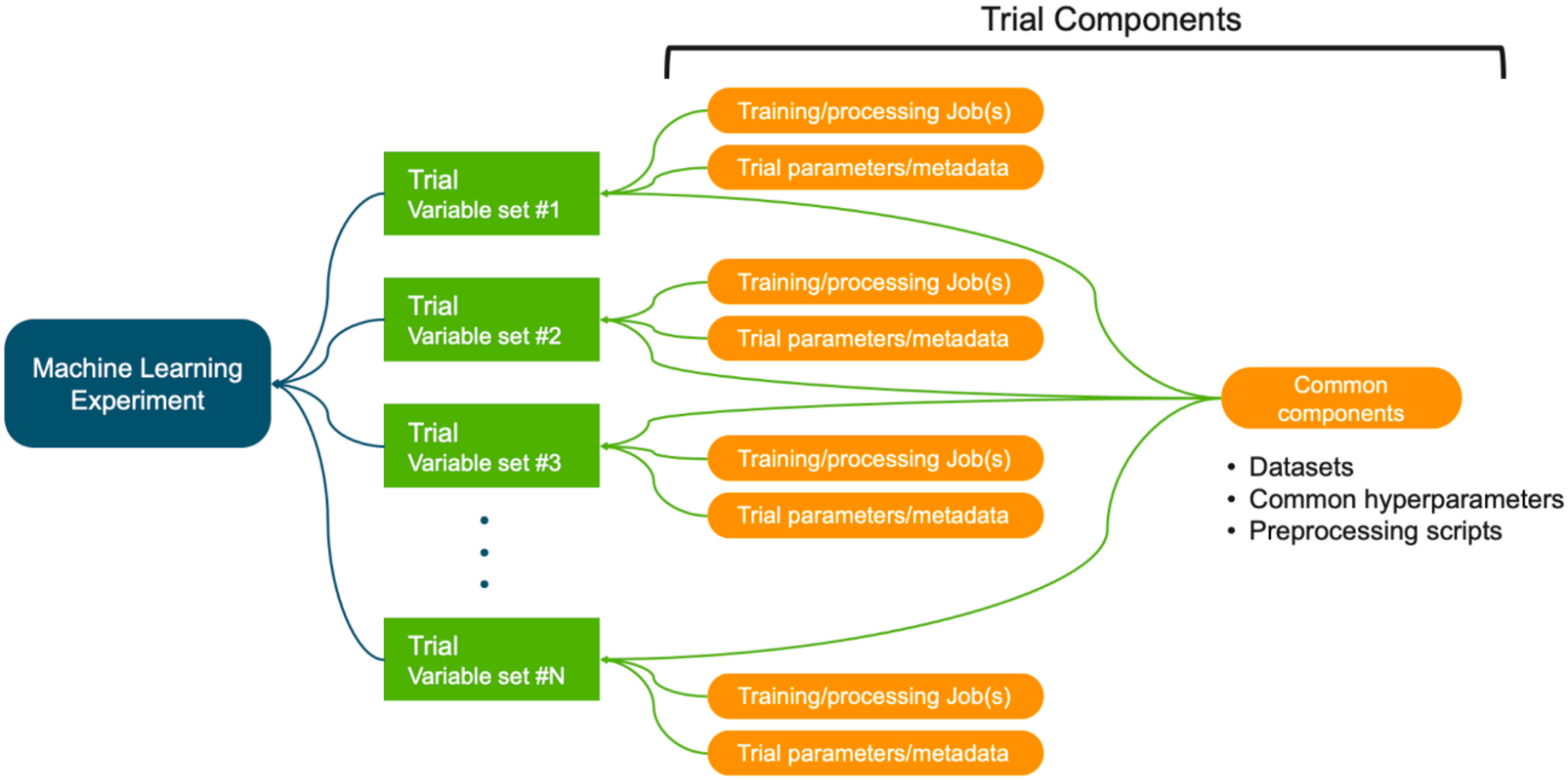
- Propose a hypothesis (e.g. Algorithm A has better predictive preformance than Algorithm B in this setting)
- Design experiment and run multiple trials with various variable values (e.g. Fitting and evaluating predictive preformance of Algorithm A and Algorithm B on both simulated and real datasets)
- Intrepret results to accept or reject your hypothesis (e.g. Algorithm A does achieve better predictive preformance but requires a lot more computation)

Why running a ML experiment is difficult

- Too many components to track (parameters, artifacts, jobs, design decisions, etc.)
- Too many ways results could go wrong (numerical instability, un-tuned hyperparameters, data representation, model misspecification, etc.)

A non-exhaustive list of things to keep track of:

- Parameters: hyperparaemers, model class / architecture, hyperparameters, optimization procedure
- Jobs: pre-prossing job, training jobs, post-processing job, compute resources
- Artifacts: datasets, checkpoints, dependencies
- Metrics: training and evaluation accuracy, computation, speed
- Deubg data: Weights, gradients, objective value, optimizer state
- …


Trial Components
Machine Learning Experiment
Trial Variable set #1
Trial Variable set #2
Trial Variable set #3
Trial Variable set #N
Training/processing Job(s)
Trial parameters/metadata
Training/processing Job(s)
Trial parameters/metadata
Training/processing Job(s)
Trial parameters/metadata
Training/processing Job(s)
Trial parameters/metadata
Common components

- Datasets
- Common hyperparameters
- Preprocessing scripts

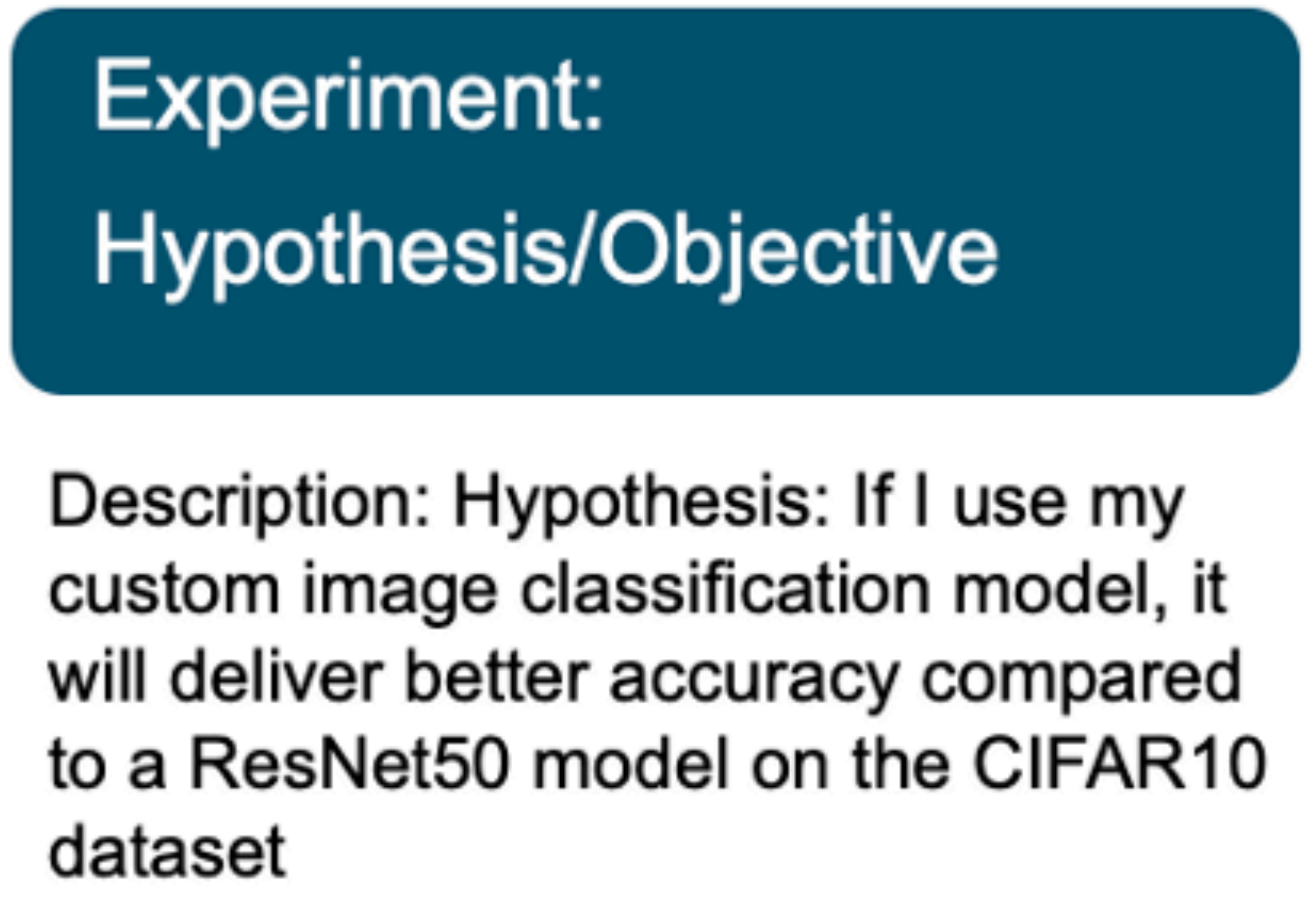
Example variable sets:
{'optimizer': 'adam', 'model': 'resnet', 'epochs': 30}
{'optimizer': 'sgd', 'model': 'custom', 'epochs': 120}

Source: Towards Datascience Blog: https://towardsdatascience.com/a-quick-guide-to-managing-machine-learning-

Step 1: Formulate a hypothesis and design an experiment

- Experiment is uniquely defined by an objective or hypothesis

- Experiments should contain more than one trial (make a convincing case that results are valid and stable to perturbations)

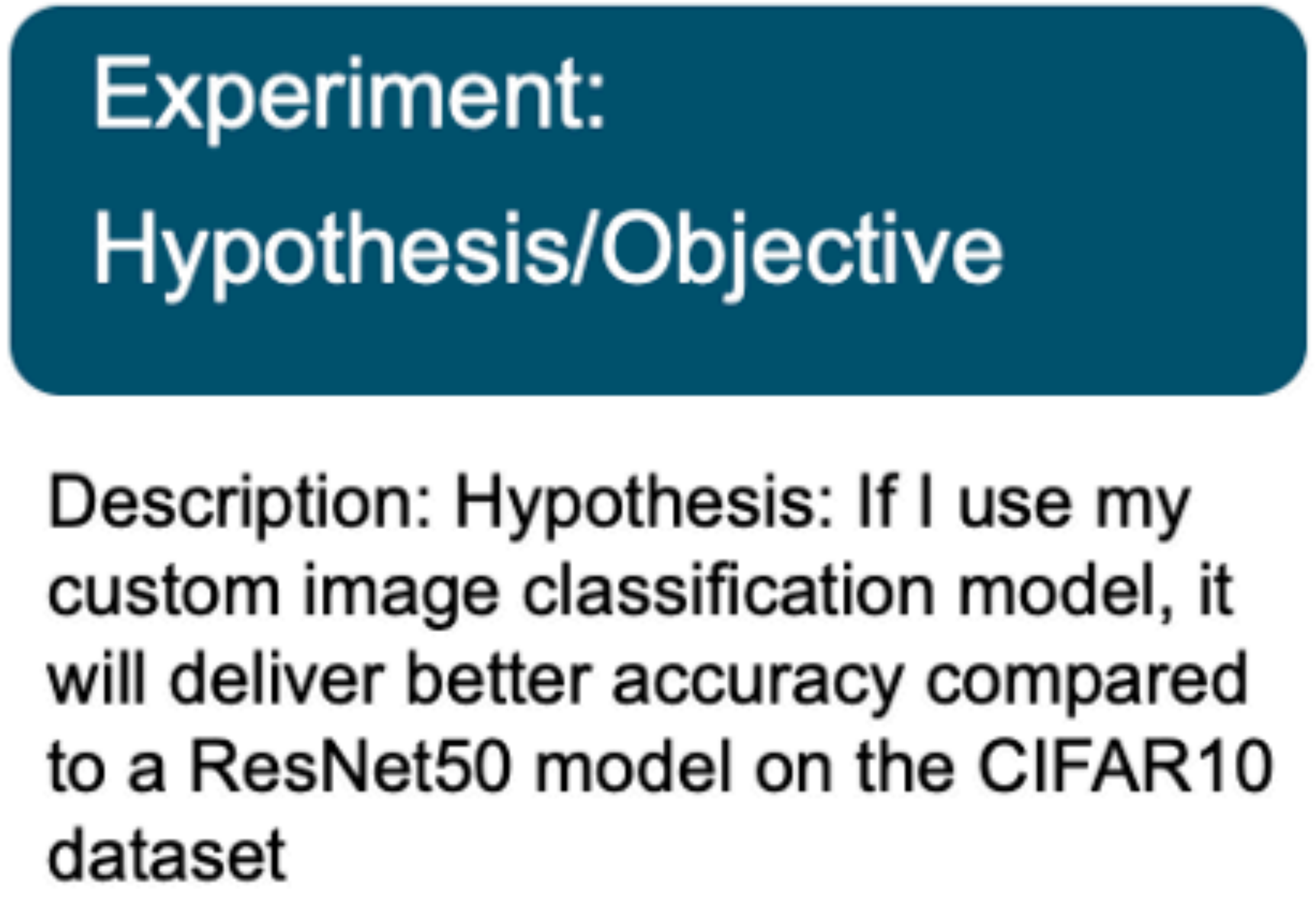


Experiment:

Hypothesis/Objective

Description: Hypothesis: If I use my custom image classification model, it will deliver better accuracy compared to a ResNet50 model on the CIFAR10 dataset

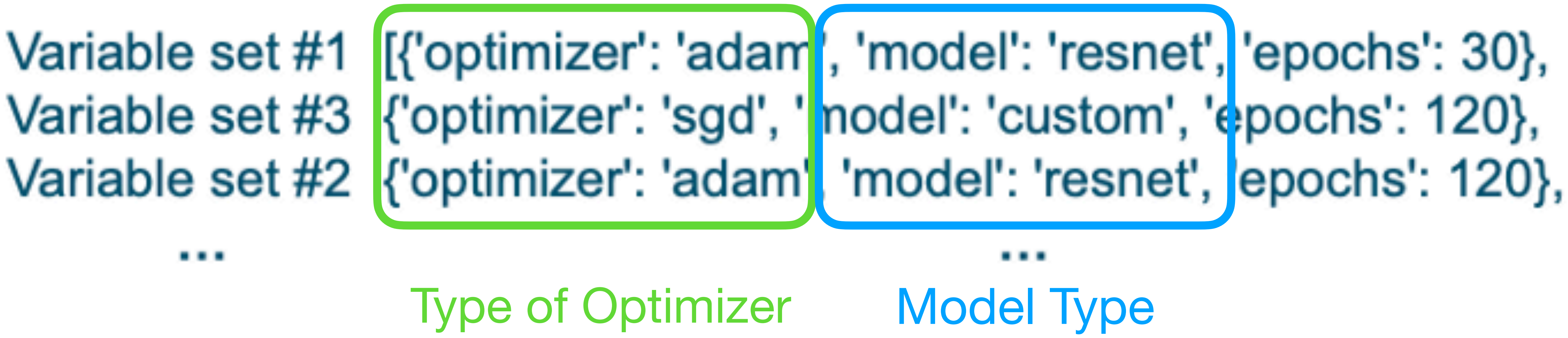
Step 2: Define experiment variables



Experiment:

Hypothesis/Objective

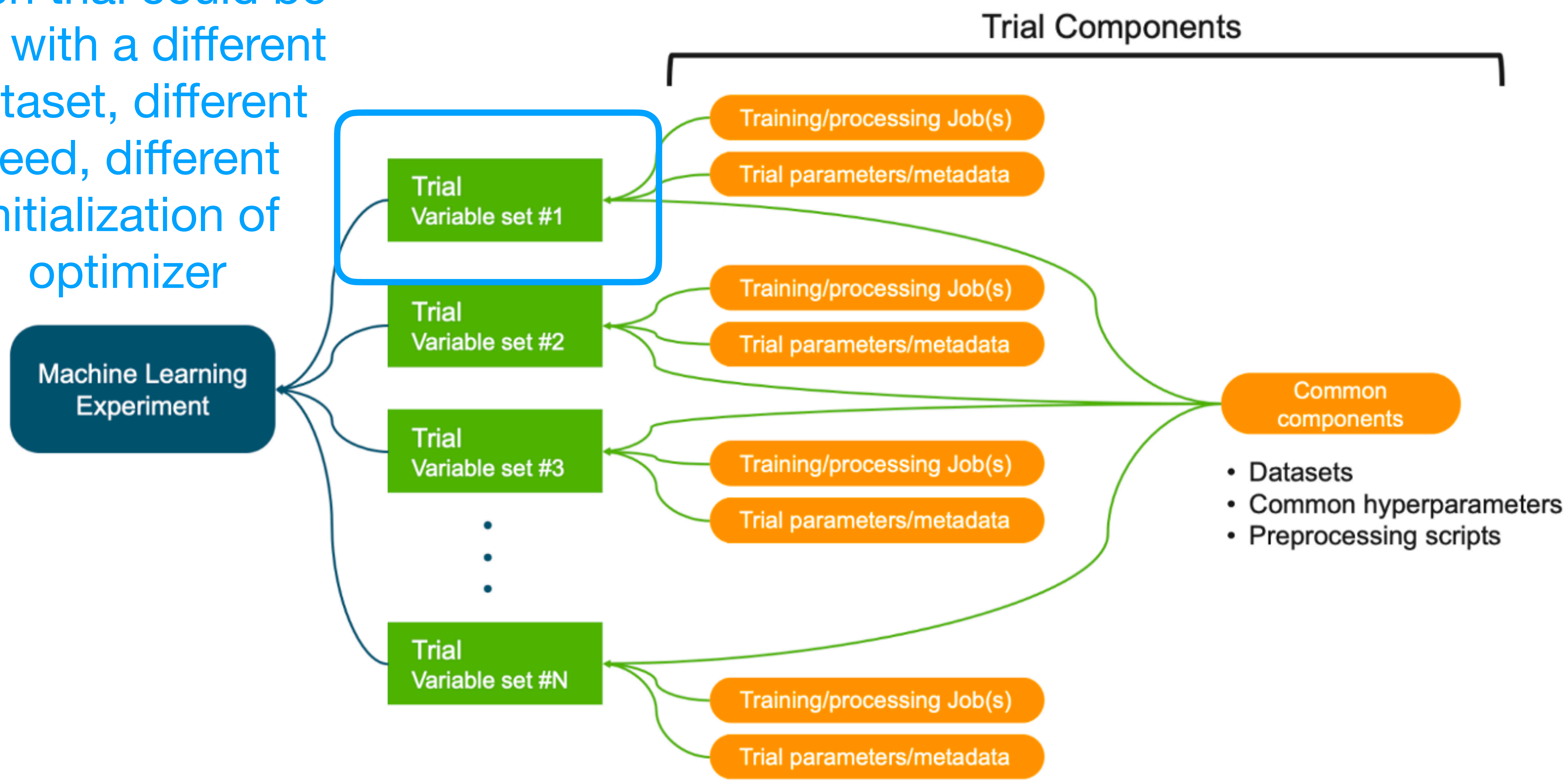
Description: Hypothesis: If I use my custom image classification model, it will deliver better accuracy compared to a ResNet50 model on the CIFAR10 dataset



```
Variable set #1  [{'optimizer': 'adam', 'model': 'resnet', 'epochs': 30},
Variable set #3  {'optimizer': 'sgd', 'model': 'custom', 'epochs': 120},
Variable set #2  {'optimizer': 'adam', 'model': 'resnet', 'epochs': 120},
        ...                                  ...
```

Step 3: Run Trials and Jobs

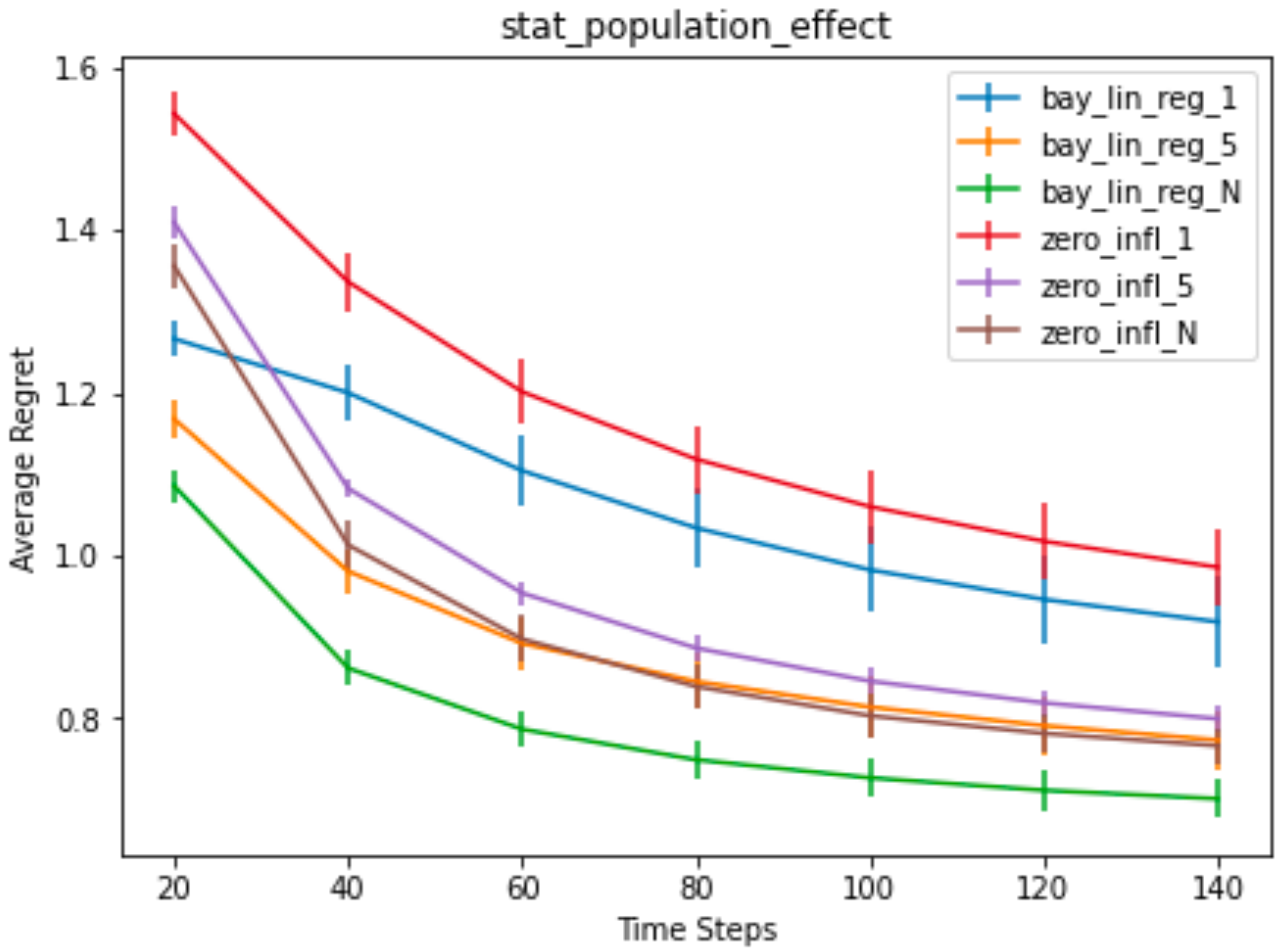
Each trial could be run with a different dataset, different seed, different initialization of optimizer


Trial Components
Machine Learning Experiment
Trial Variable set #1
Trial Variable set #2
Trial Variable set #3
Trial Variable set #N
Training/processing Job(s)
Trial parameters/metadata
Training/processing Job(s)
Trial parameters/metadata
Training/processing Job(s)
Trial parameters/metadata
Training/processing Job(s)
Trial parameters/metadata
Common components

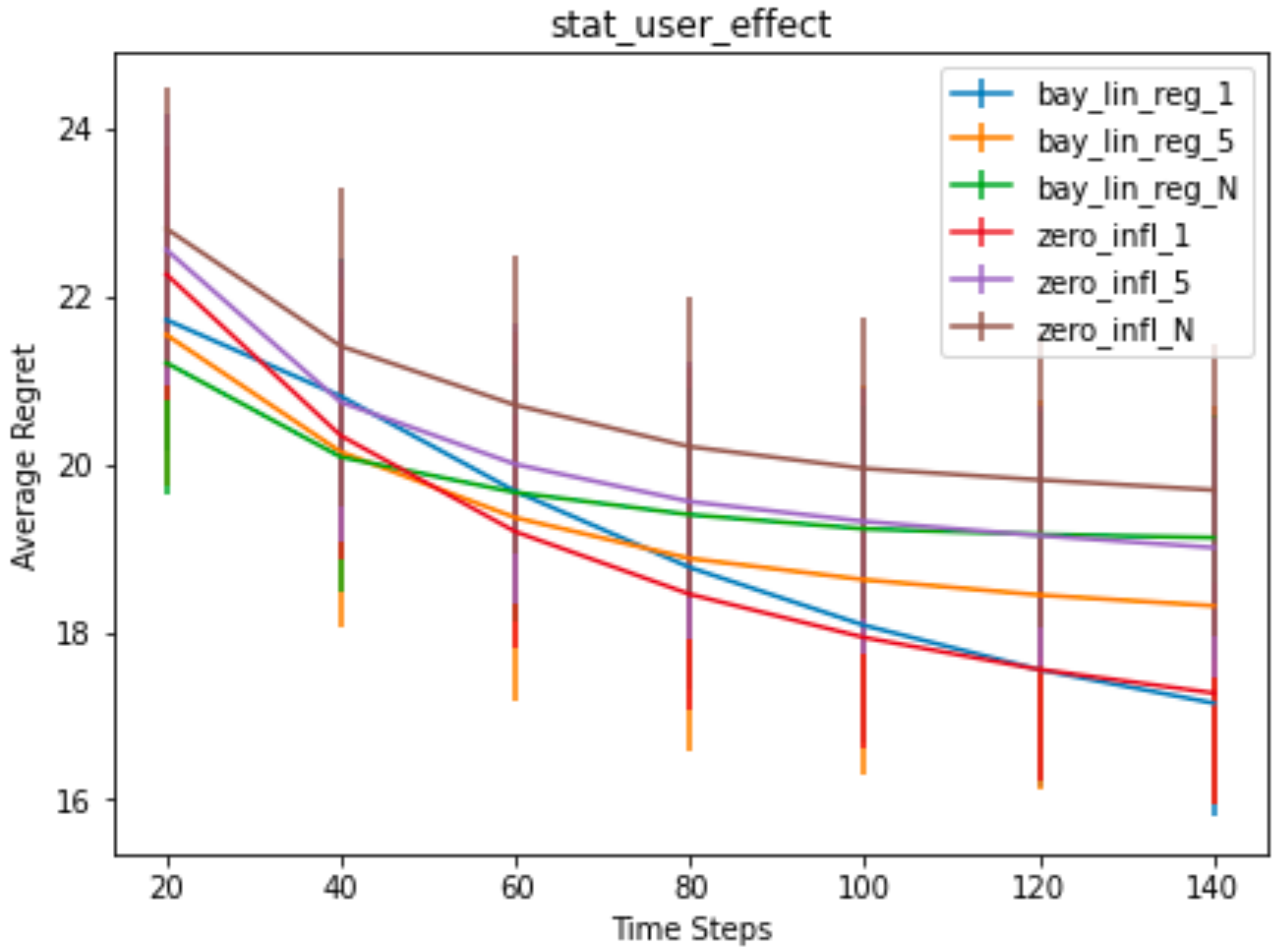
- Datasets
- Common hyperparameters
- Preprocessing scripts

Example variable sets:
{'optimizer': 'adam', 'model': 'resnet', 'epochs': 30}
{'optimizer': 'sgd', 'model': 'custom', 'epochs': 120}

Step 4: Interpret Results



stat_population_effect



stat_user_effect

A Non-Exhaustive List of Tips

Pre-processing Tips

- Dataset splitting
  - e.g. Bootstrap or CV
- Data cleaning and formatting
  - creating state space, feature selection, PCA, etc.
  - normalize values close to [-1, 1] for numerical stability
- Model selection
  - Choosing model candidates
- Optimizer selection
  - Choosing a method of fitting the model

Experiment Tips

- Testing the validity of your objective
  - Generating your own toy data with ground truth. Does your objective assign high likelihood or low error to the ground truth over other values?
- Stability and Reproducibilty
  - Seeding your trials
  - If you slightly perturb the data, how much do your results differ from each other?
- Hand Code vs. Built in Packages
  - Understand the trade off between investing in a hand coded method vs. using a built in package
- Debugging: Isolate the Issue
  - Is it the data? Is it the optimizer? Is it the hyperparameters? Is it my model?

Post-Processing Tips

- Graphs, Figures, Tables
  - What metrics are helpful to report?
- Save experiment values first and then generate figures!
- Interpretation of Results
  - Do the results make sense? Is this because of a bug or because of an assumption?