End-to-end Deep Learning of Optimization Heuristics

http://chriscummins.cc/pact17


Chris Cummins University of Edinburgh


Pavlos Petoumenos University of Edinburgh


Zheng Wang Lancaster University


Hugh Leather University of Edinburgh

THE UNIVERSITY of EDINBURGH informatics

EPSRC Centre for Doctoral Training in Pervasive Parallelism

EPSRC Engineering and Physical Sciences Research Council

compilers are <u>very</u> complex

```
int main(
 int argc,
 char** arg)
 {...
```

{hundreds, thousands, **millions**} of choices

```
_main:
   .cfi_start
proc
## BB#0:
   pushq %rbp
...
```

hand-coded heuristics
(out of date by time of release)

Machine learning in compilers

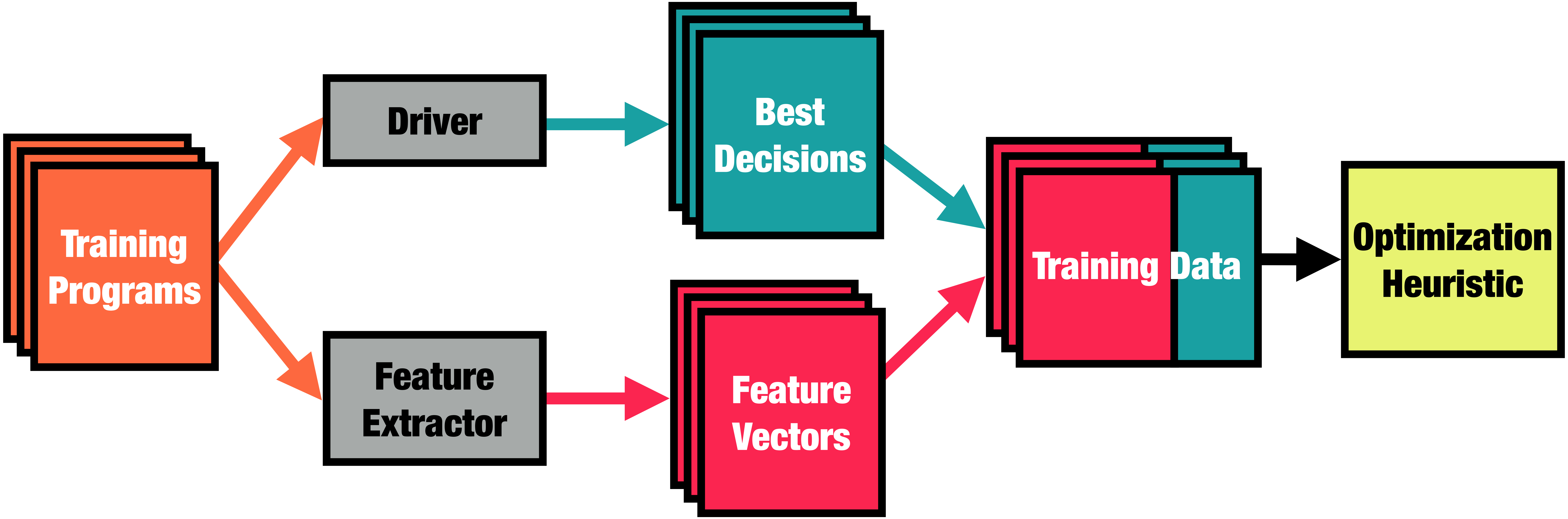
optimization decision

features (derived from IR)

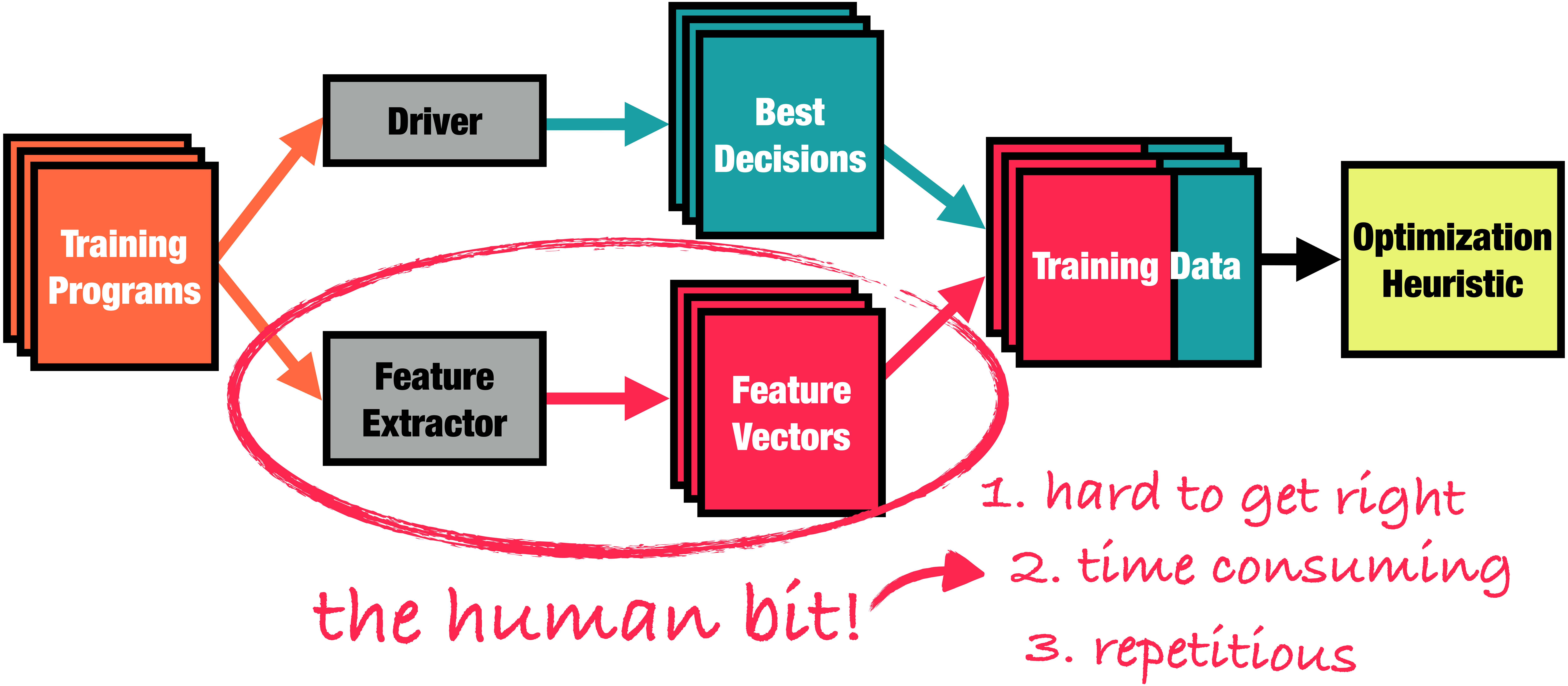
$$y = f(x)$$

model

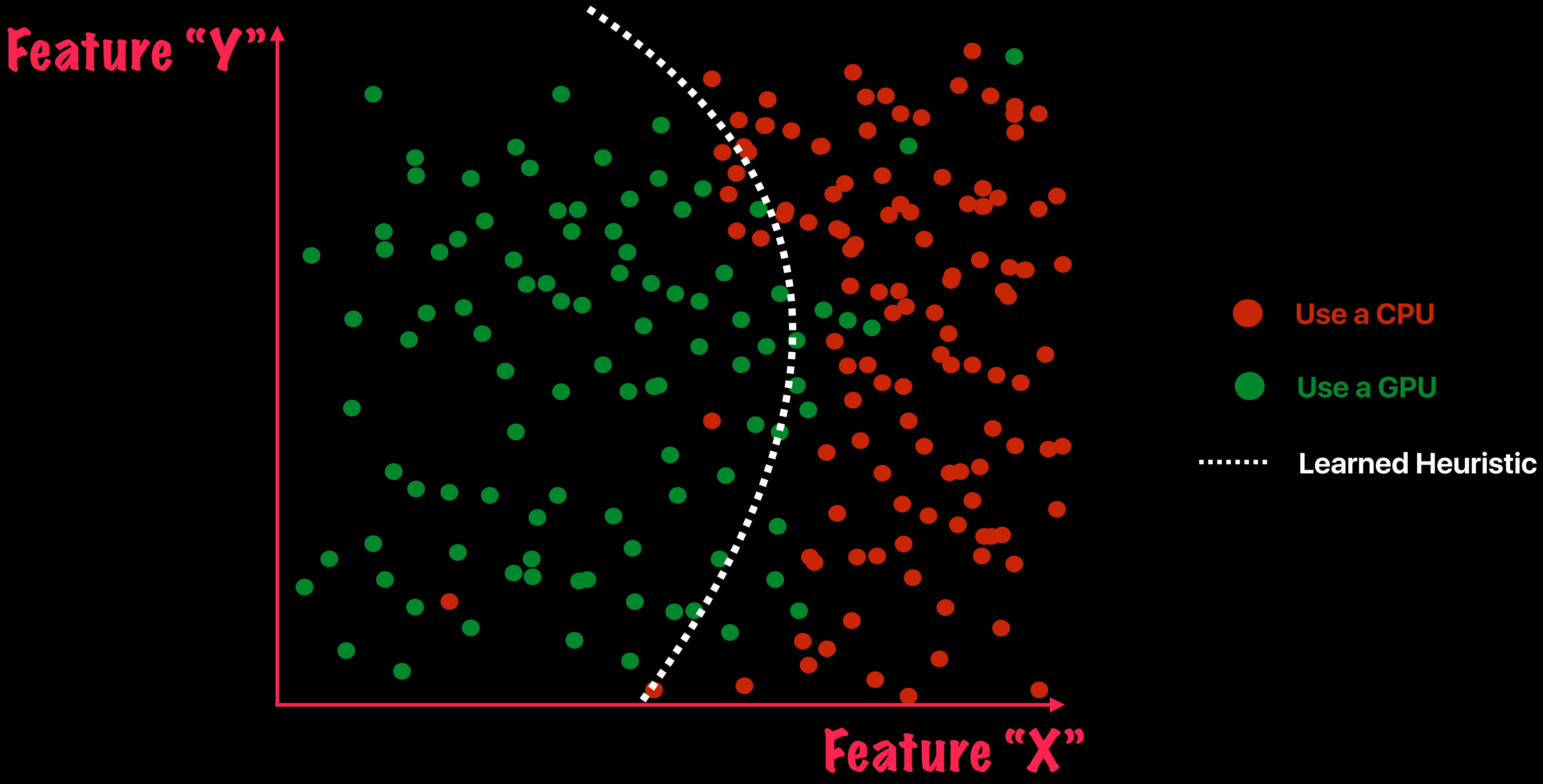
Machine learning in compilers



Machine learning in compilers



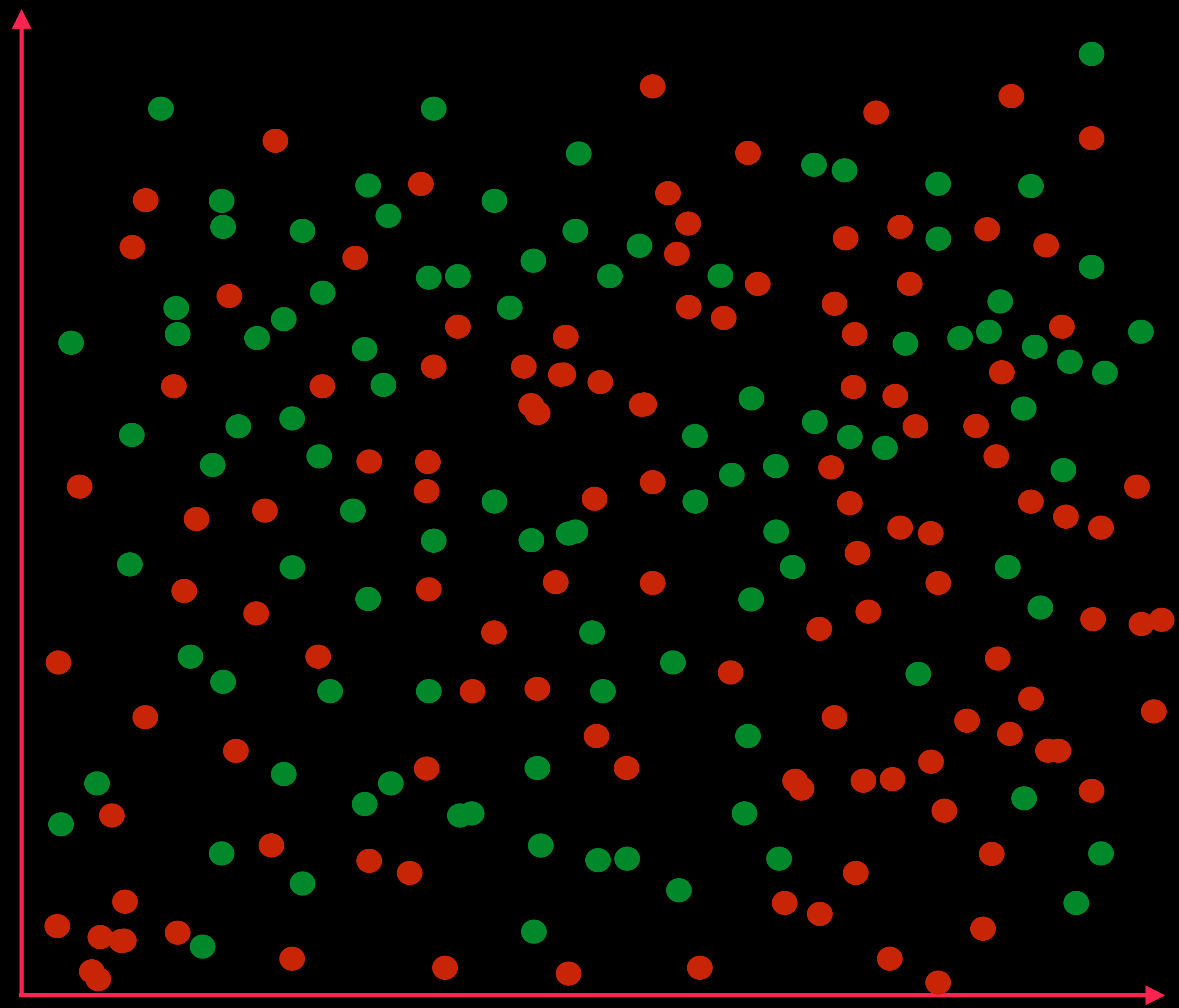
Feature space



Feature space

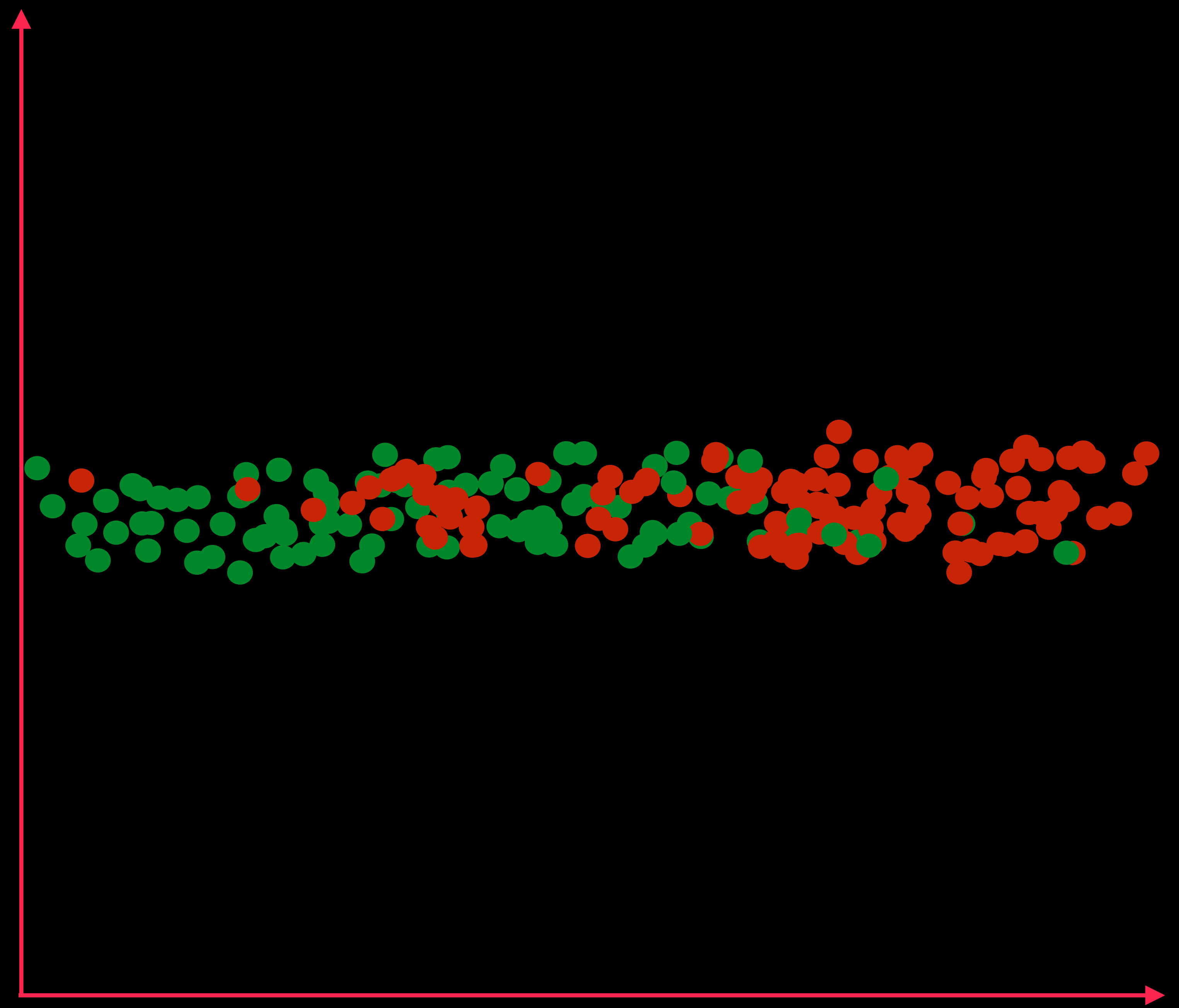
need good features!

Ways to fail



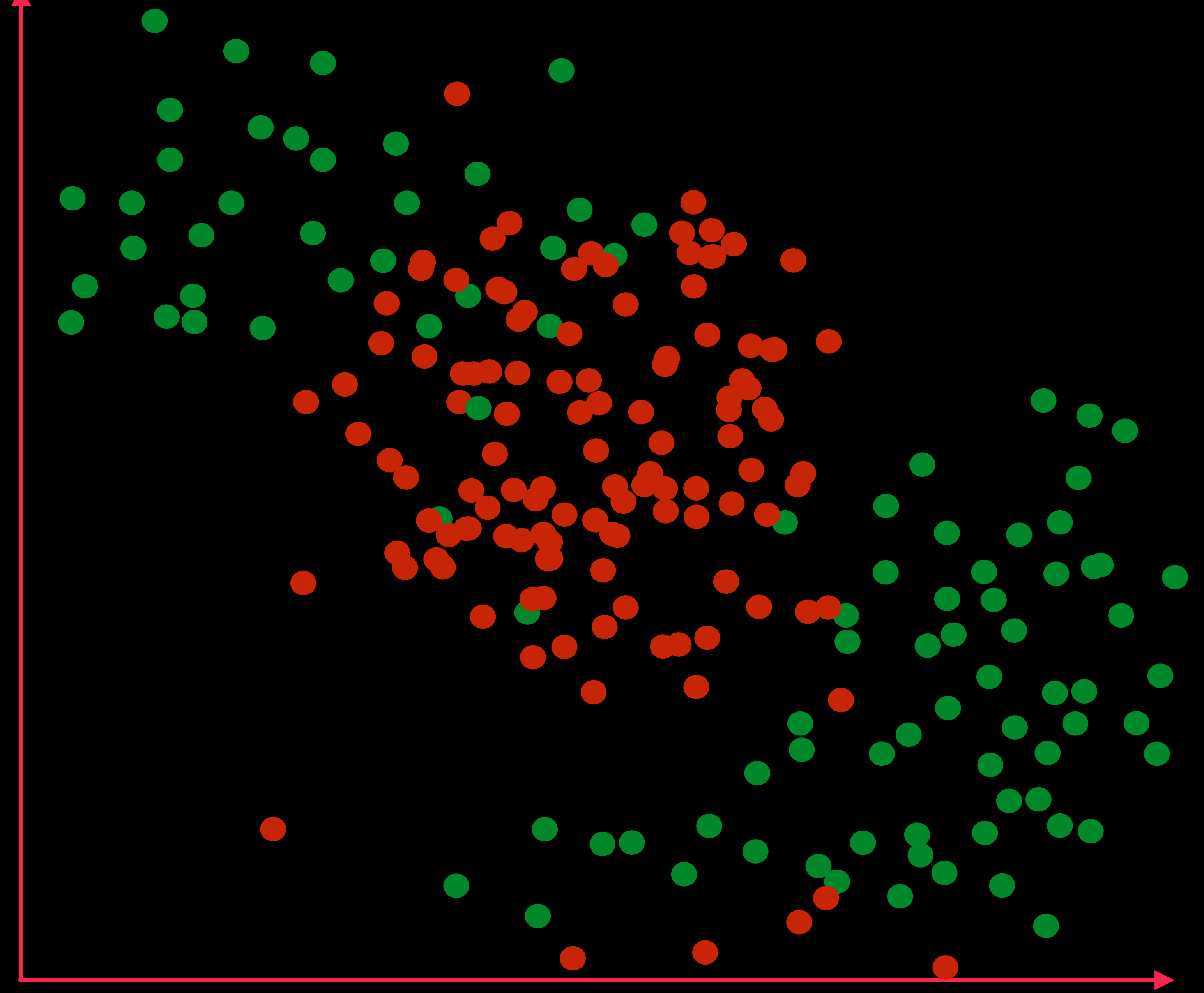
irrelevant

e.g. not capturing the right information



incomplete

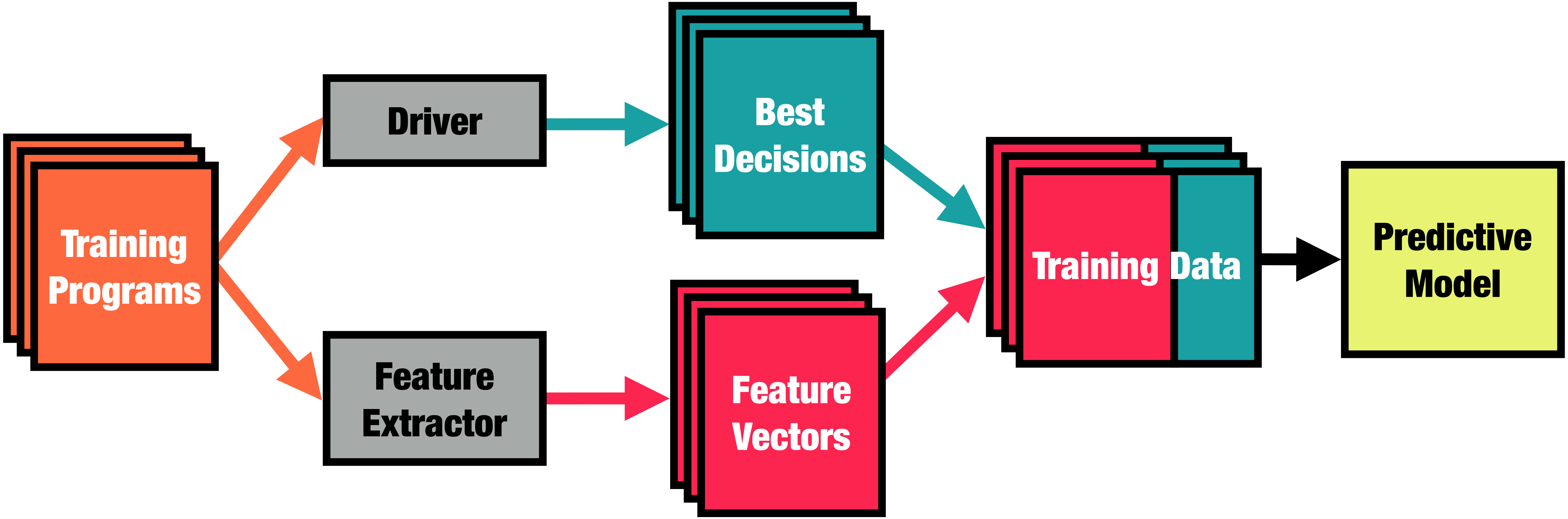
e.g. missing critical information



unsuitable

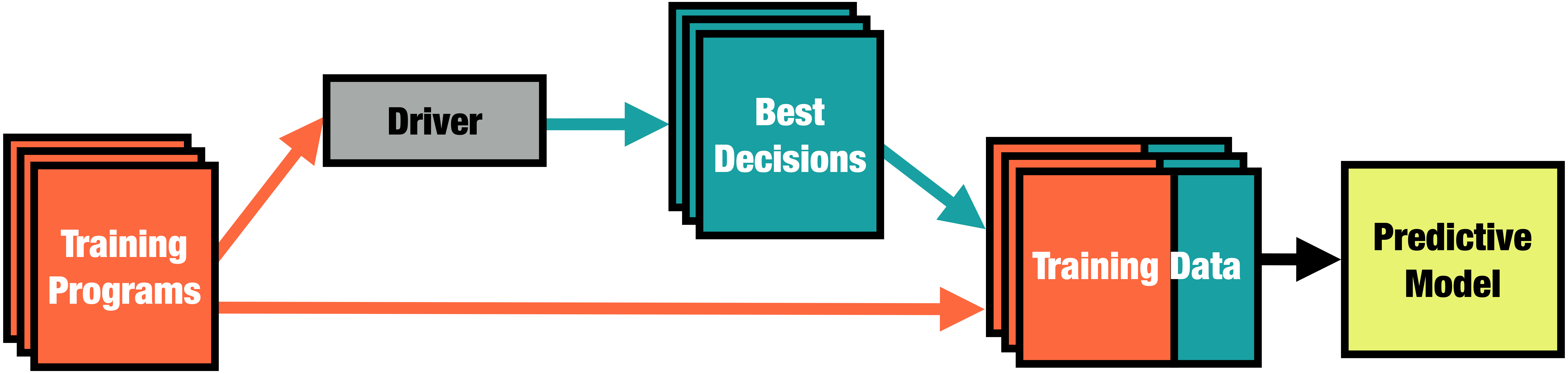
e.g. wrong combination of features / model

What we have



Training Programs
Driver
Feature Extractor
Best Decisions
Feature Vectors
Training Data
Predictive Model

What we need


Driver
Best
Decisions
Training
Programs
Training Data
Predictive
Model

Contributions

Heuristics <u>without</u> features

<u>Beats</u> expert approach

Learning <u>across</u> heuristics

Our approach

Program Code

Deep Learning

Optimization Decision

Our approach

Our approach

Our approach

How does it work?

kernel void memset_kernel(
global char* mem_d, short val...
kernel void A(
global char* a, short b...
Code in
Rewriter
Encoder
Vocabulary
␣ 0
(1
* 2
, 3
A 4
a 5
b 6
char 7
global 8
kernel 9
short 10
void 11
Embed-
ding
Language
Model
Heuristic
Model

How well does it work?

Prior Art

Heterogeneous Mapping


Portable Mapping of Data Parallel Programs to OpenCL for Heterogeneous Systems

Dominik Grewe Zheng Wang Michael F.P. O'Boyle

School of Informatics, University of Edinburgh

{dominik.grewe, zh.wang}@ed.ac.uk, mob@inf.ed.ac.uk

Abstract

1. Introduction

CGO'13

Grewe et. al

Thread Coarsening


Automatic Optimization of Thread-Coarsening for Graphics Processors

Alberto Magni, School of Informatics, University of Edinburgh, United Kingdom, a.magni@sms.ed.ac.uk

Christophe Dubach, School of Informatics, University of Edinburgh, United Kingdom, christophe.dubach@ed.ac.uk

Michael O'Boyle, School of Informatics, University of Edinburgh, United Kingdom, mob@inf.ed.ac.uk

ABSTRACT

1. INTRODUCTION

PACT'14

Magni et. al

Prior Art

Heterogeneous Mapping

Thread Coarsening

Decision Space

Binary classification
{CPU, GPU}

One-of-six classification
{1, 2, 4, 8, 16, 32}

Model

Decision Tree

Cascading
Neural Networks

CGO'13

PACT'14

Prior Art

Heterogeneous Mapping

Thread Coarsening

Features

4 features

7 features

Combined from 7 raw values.

Principle Components of 34 raw values.

2 papers!

Instruction counts / ratios.

Instruction counts / ratios / relative deltas.

CGO'13

PACT'14

Our Approach

Heterogeneous Mapping

Thread Coarsening

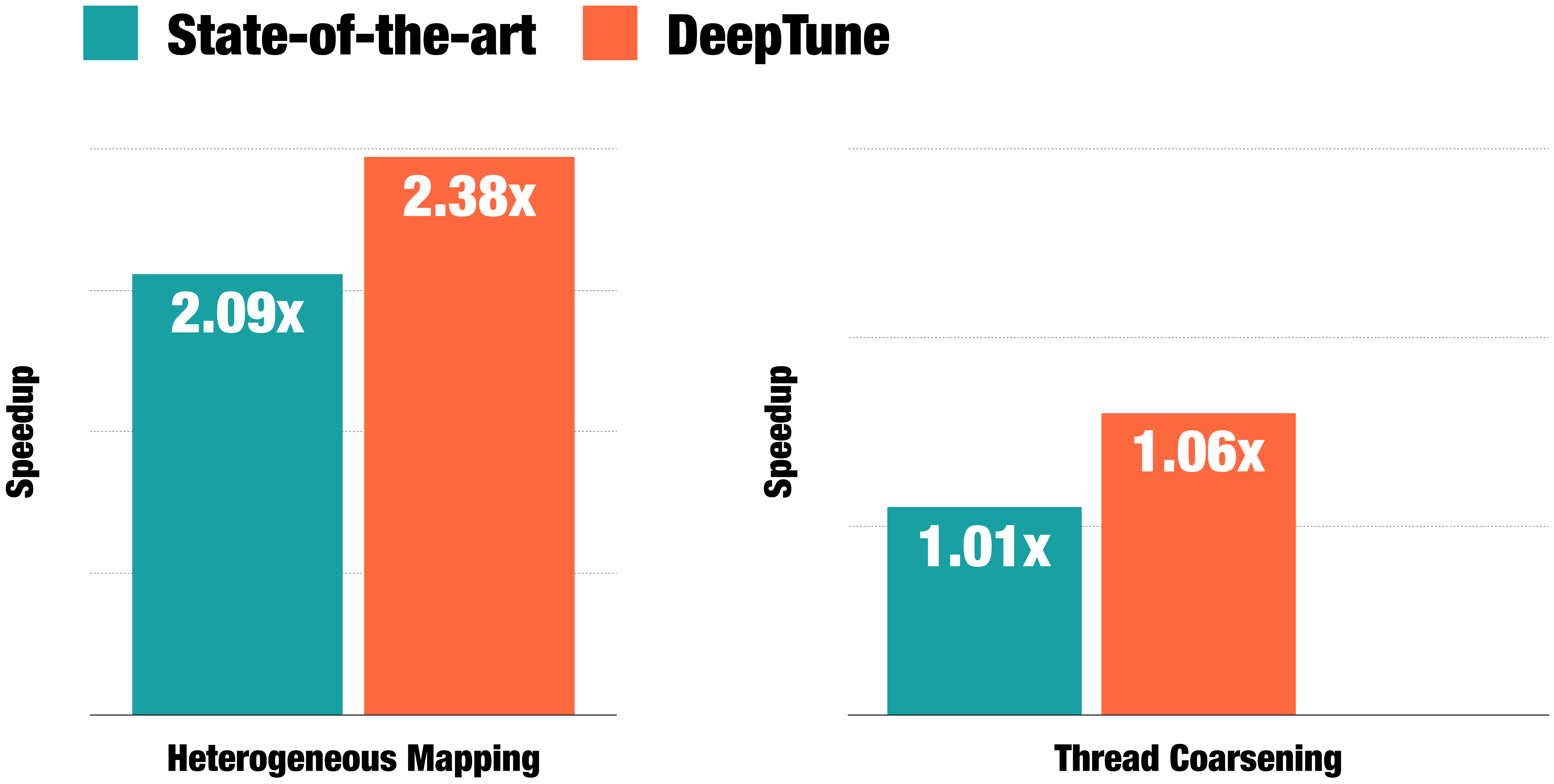
1. Use the same model design for both
2. No tweaking of parameters
3. Minimum change - 3 line diff

Prior Art

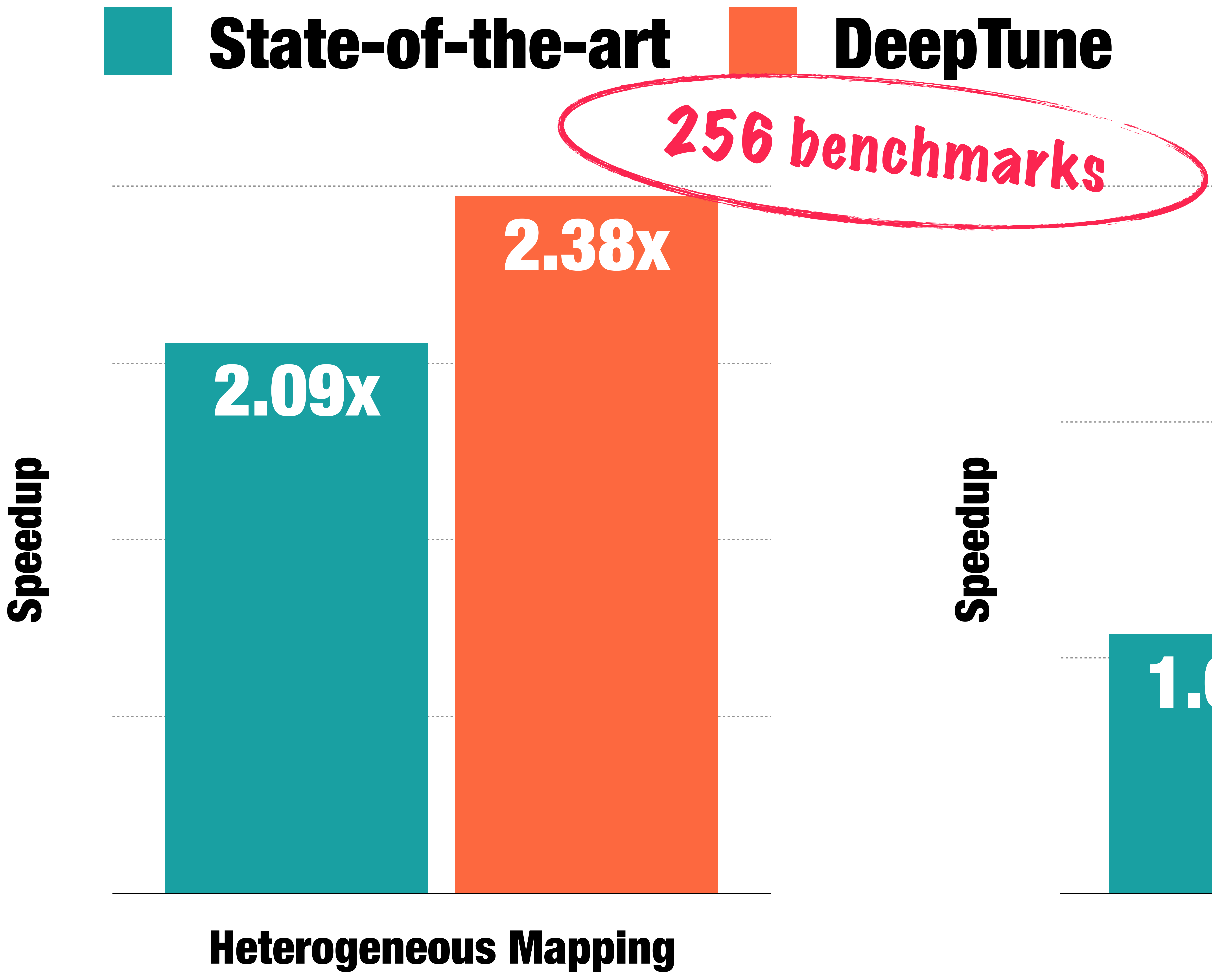
Heterogeneous Mapping	Thread Coarsening
Hardware	
2x CPU-GPU architectures	4x GPU architectures
Training Programs	
7 Benchmark Suites	3 Benchmark Suites
CGO'13	PACT'14

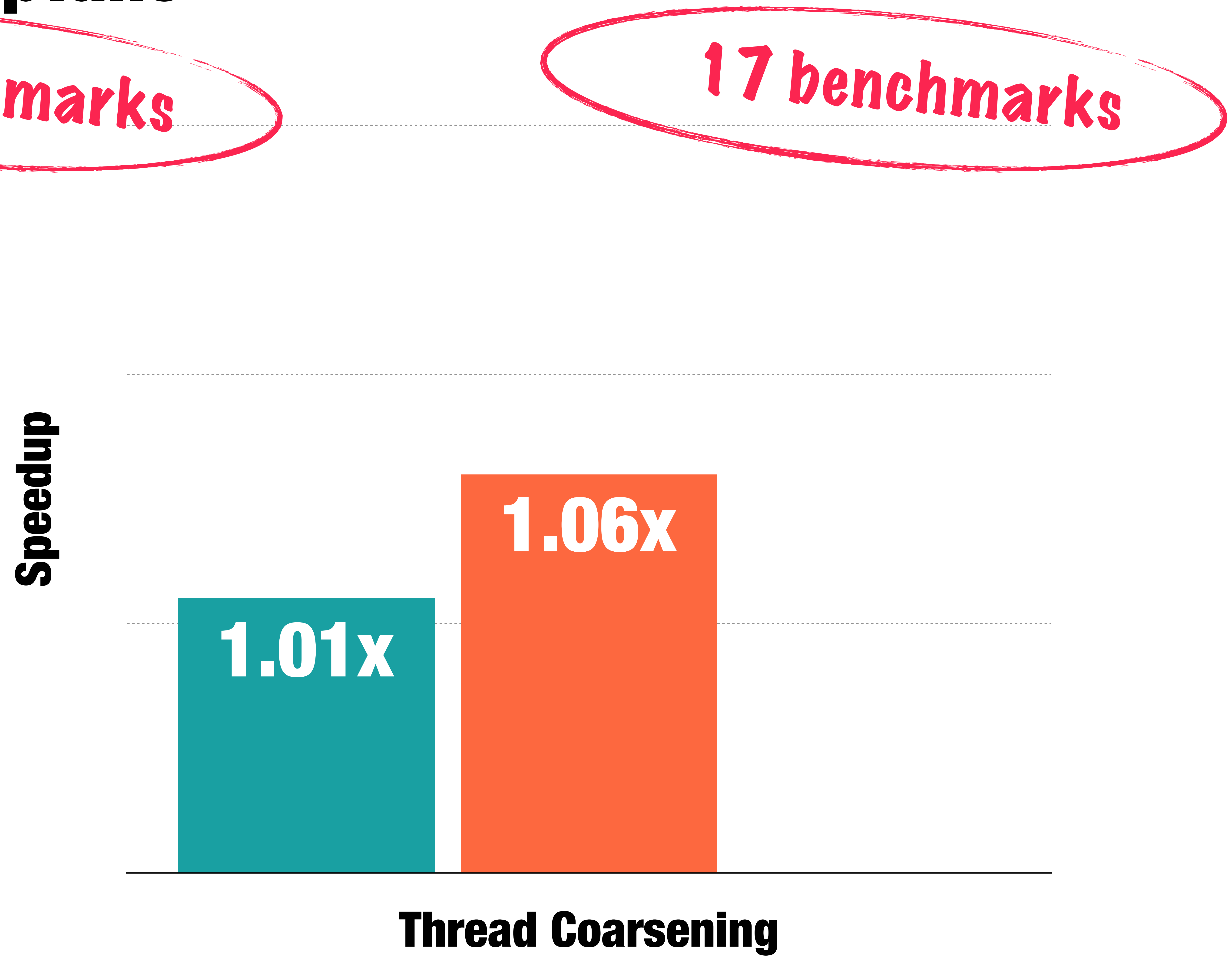
results

14% and 5% improvements over state-of-the-art



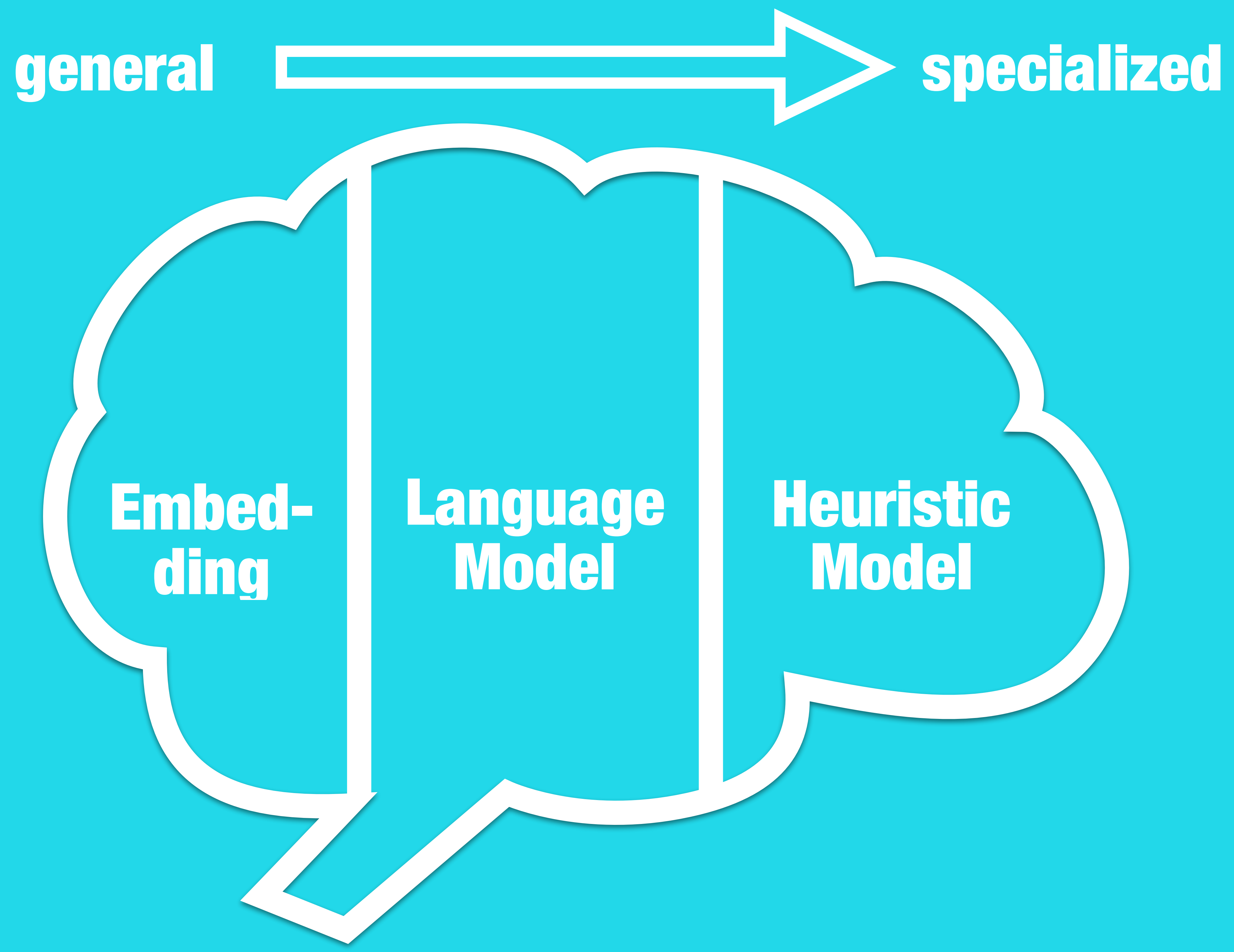
14% and 5% improvements over state-of-the-art



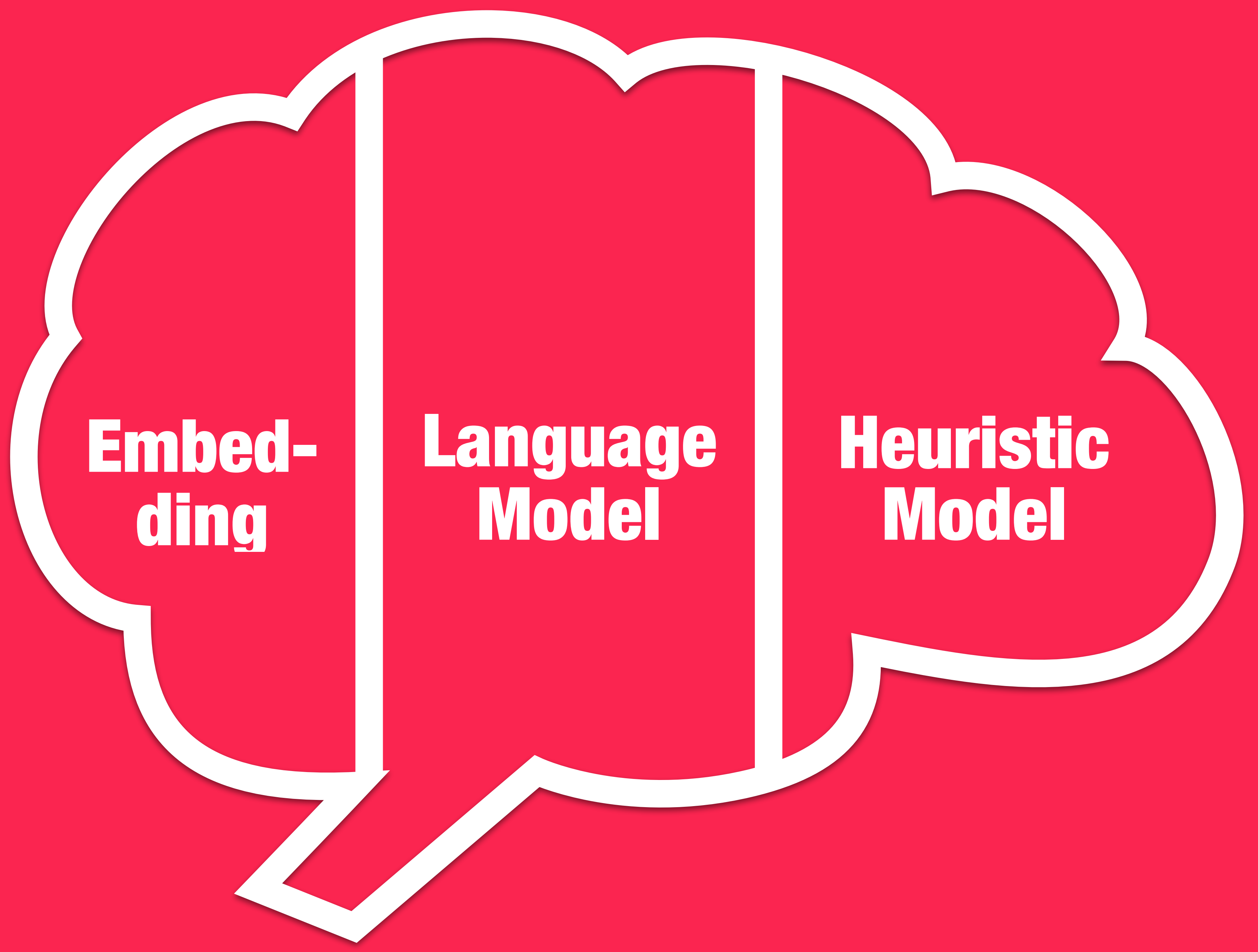


Transfer Learning

Heterogeneous Mapping



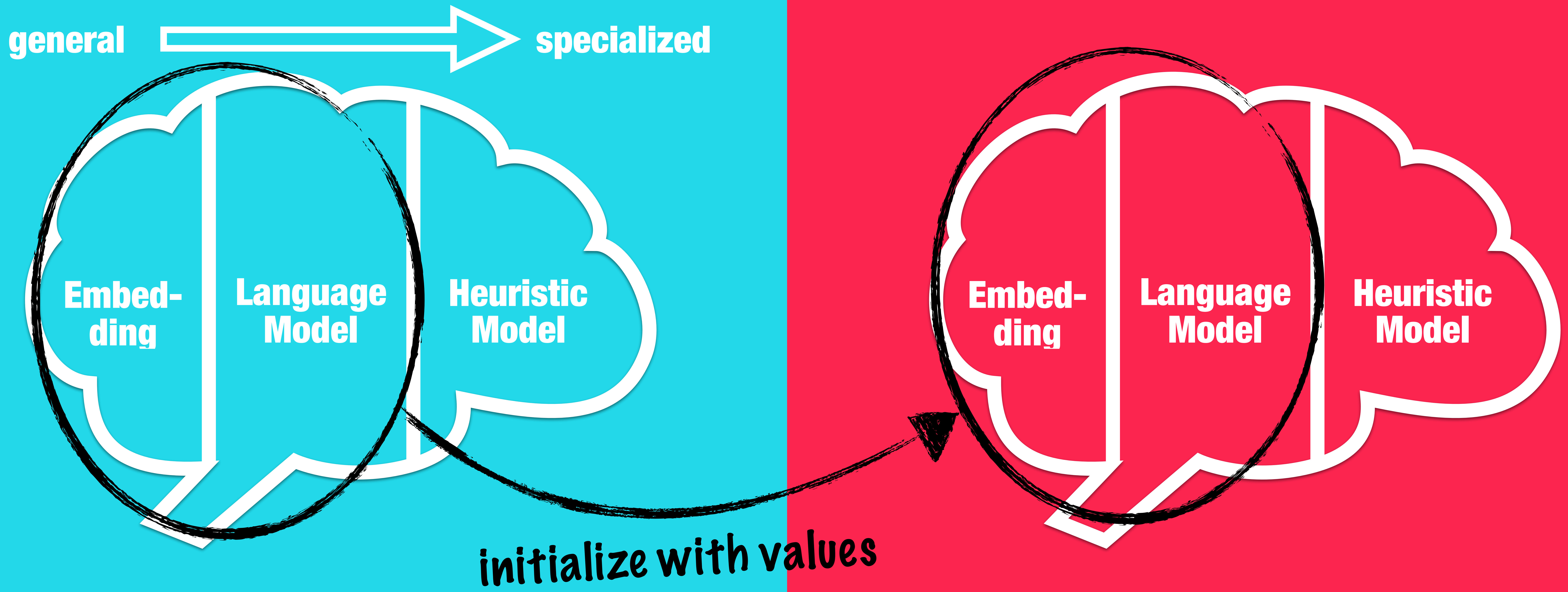
Thread Coarsening



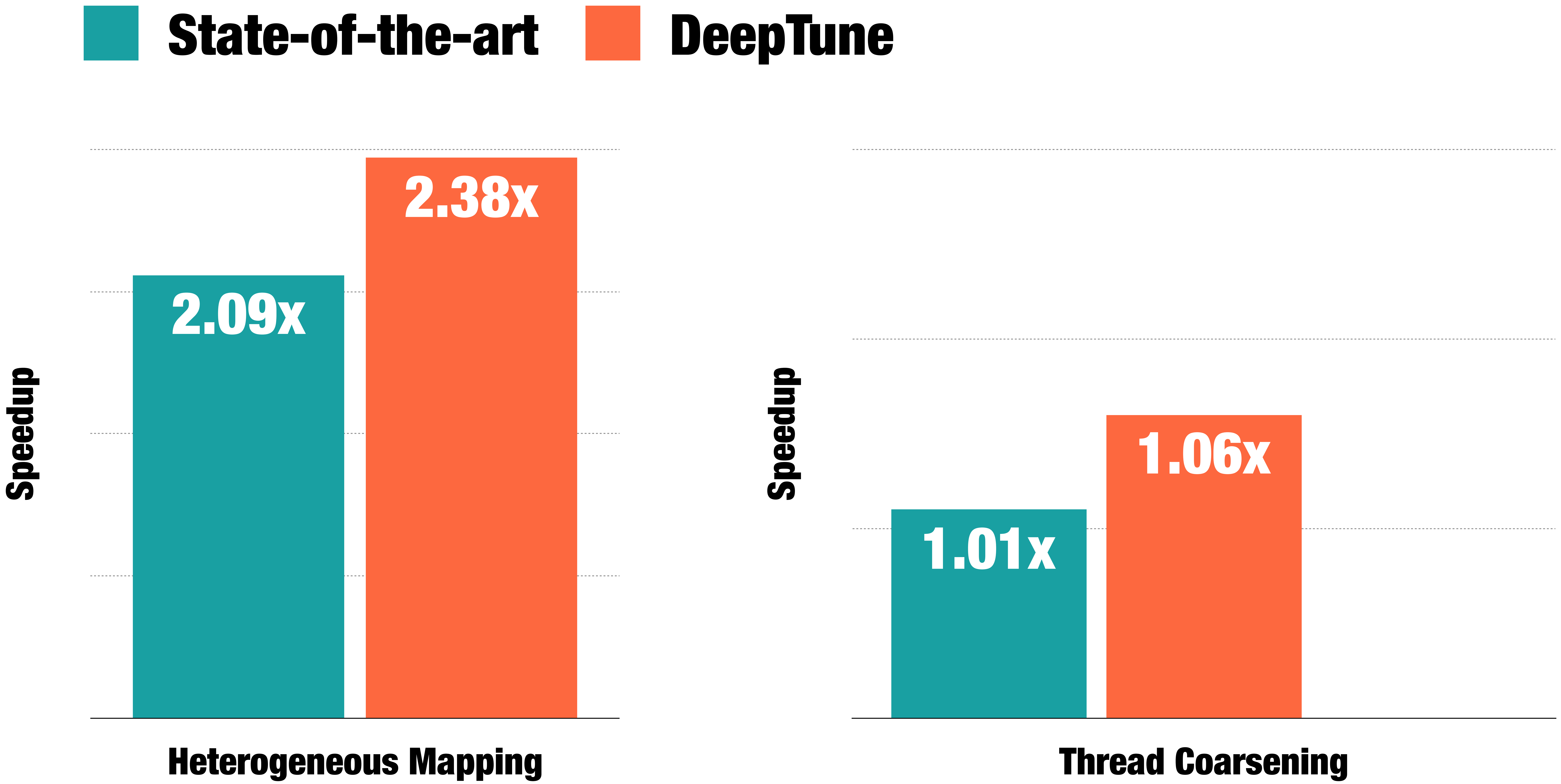
Transfer Learning

Heterogeneous Mapping

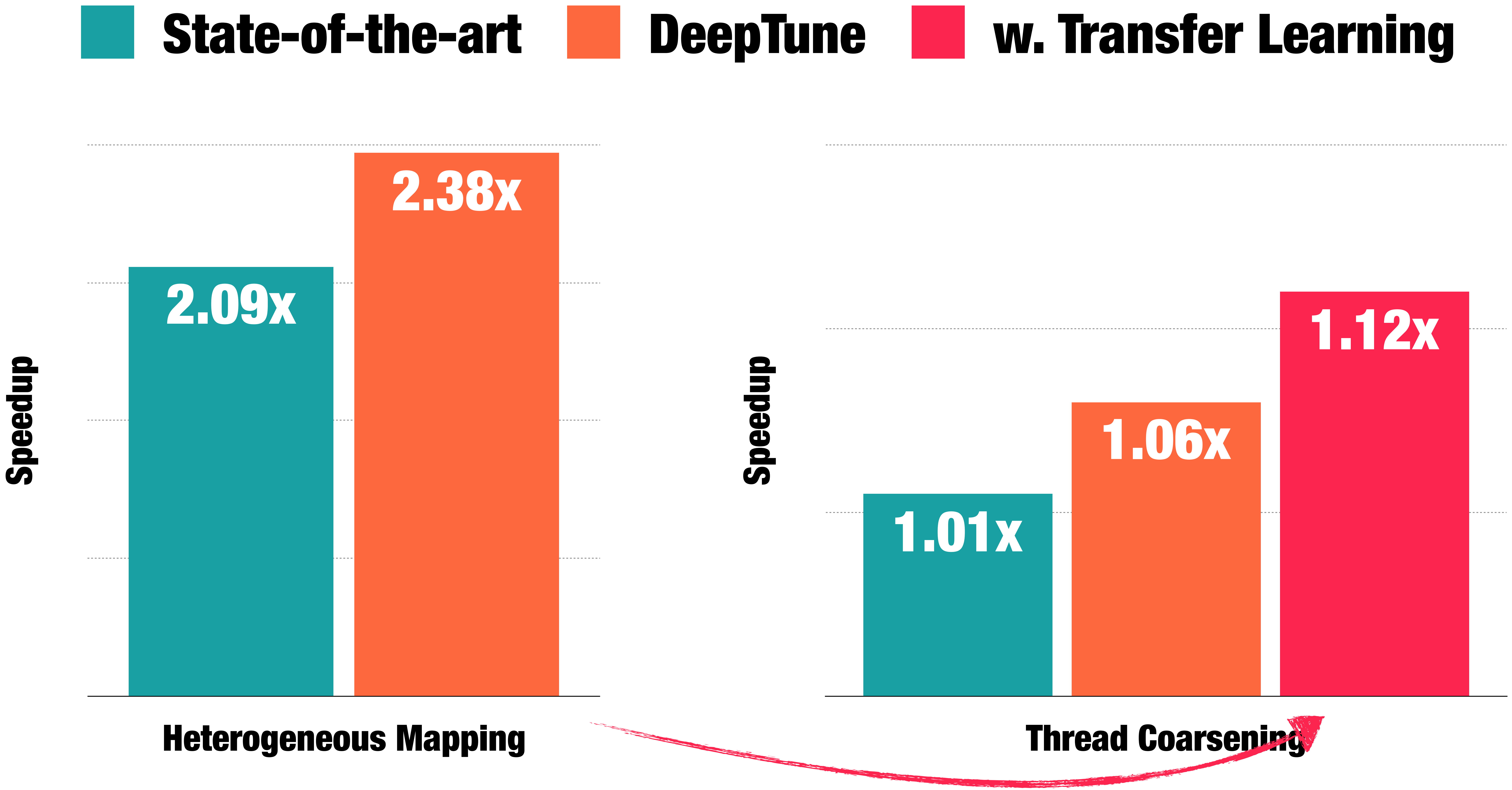
Thread Coarsening



14% and 5% improvements over state-of-the-art

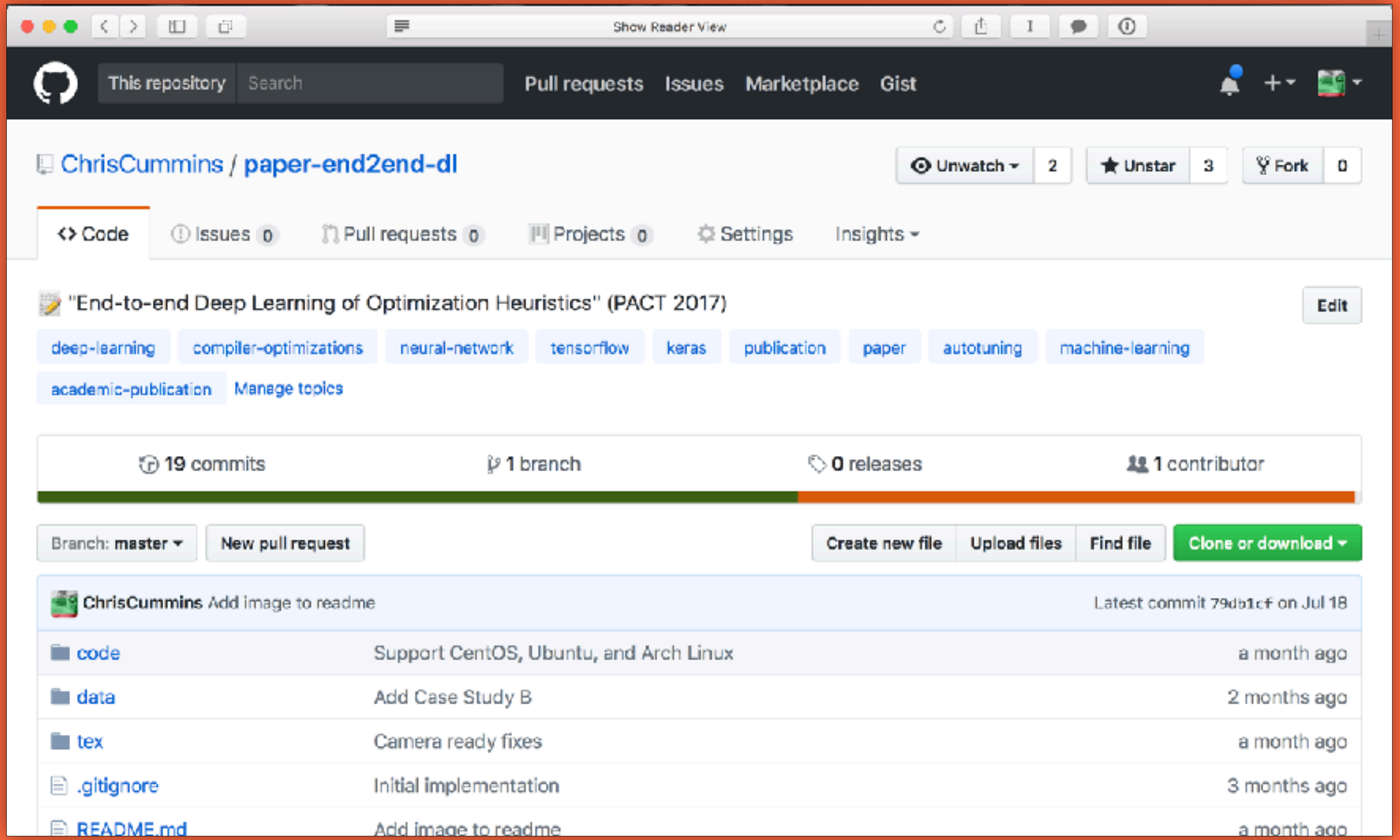


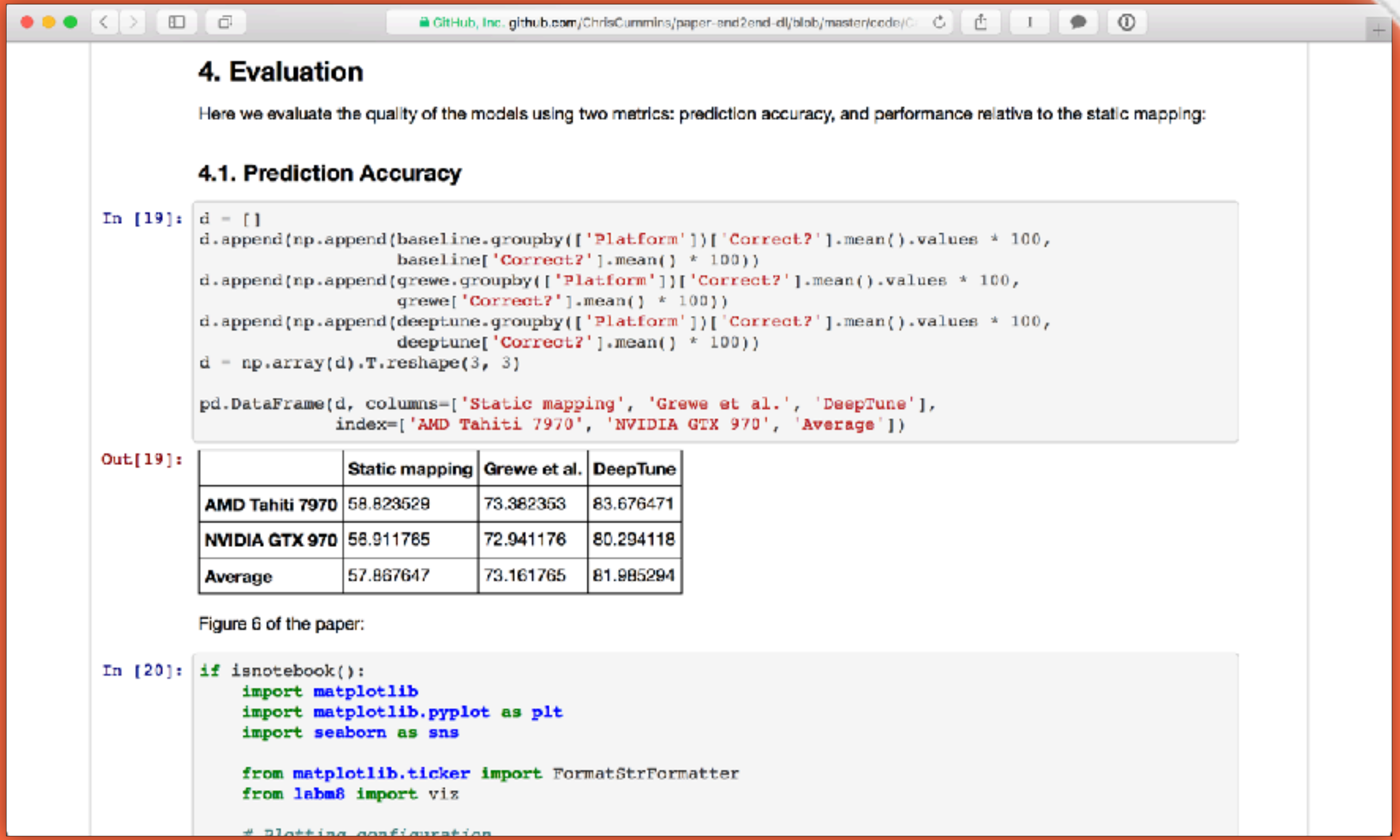
14% and 11% improvements over state-of-the-art



Try it for yourself!

Fork me on GitHub





code and data on GitHub

runs in the browser

http://chriscummins.cc/pact17

End-to-end Deep Learning of Optimisation Heuristics

Problem: feature design is hard

Featureless heuristics

First cross-domain learning

11-14% speedups

http://chriscummins.cc/pact17