Modeling Temporal Evolution in the Endomondo Fitness Data Set

Yashodhan Hemant Karandikar

Advisor : Prof. Julian McAuley

Motivation

- What will be my average heart-rate on the next run?
- How long will my next run take?

- How will my heart-rate change in the remaining part of this run?

Motivation

Temporal evolution across workouts

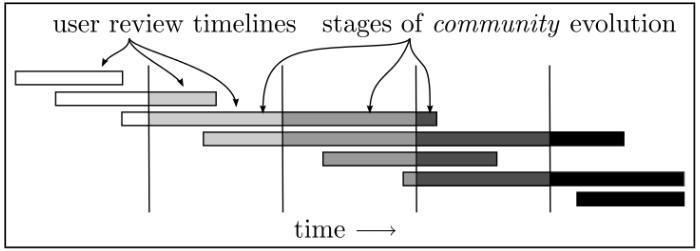
- What will be my average heart-rate on the next run?
- How long will my next run take?

Temporal evolution within workouts

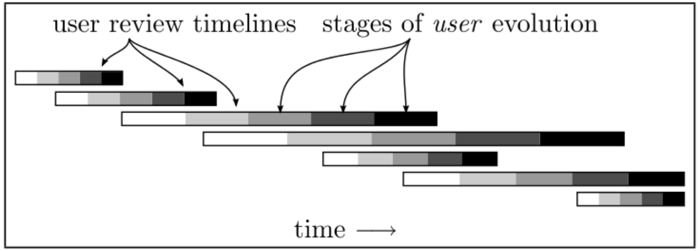
- How will my heart-rate change in the remaining part of this run?

Evolution of User Expertise in Reviews

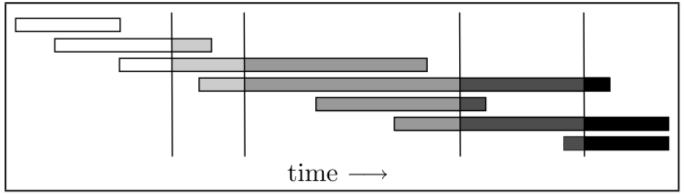
Community evolution at uniform intervals



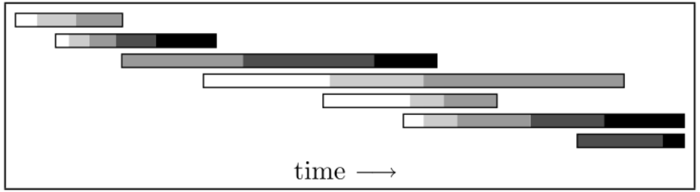
Individual user evolution at uniform intervals



Community evolution at learned intervals



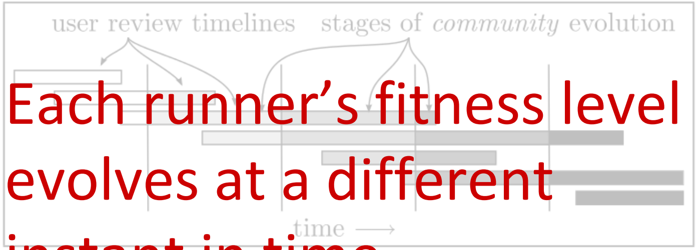
Individual user evolution at learned intervals



Source : J. J. McAuley and J. Leskovec. From amateurs to connoisseurs: modeling the evolution of user expertise through online reviews. In *World Wide Web*, 2013.

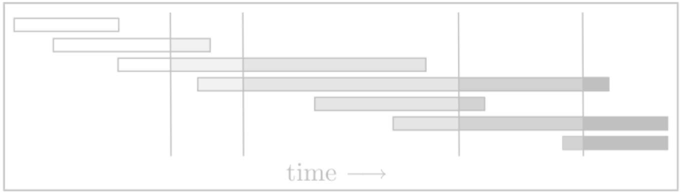
Evolution of User Expertise in Reviews

Community evolution at uniform intervals



Each runner’s fitness level evolves at a different instant in time

Community evolution at learned intervals



Individual user evolution at uniform intervals

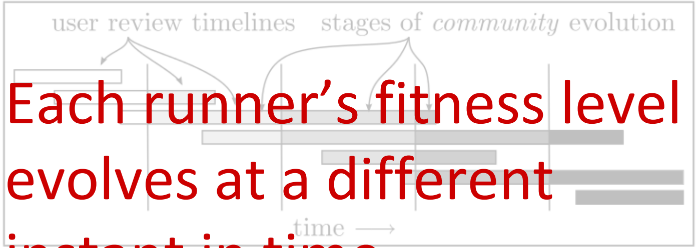
A runner’s fitness level may evolve at a non-uniform rate

Individual user evolution at learned intervals

Source : J. J. McAuley and J. Leskovec. From amateurs to connoisseurs: modeling the evolution of user expertise through online reviews. In *World Wide Web*, 2013.

Evolution of User Expertise in Reviews

Community evolution at uniform intervals



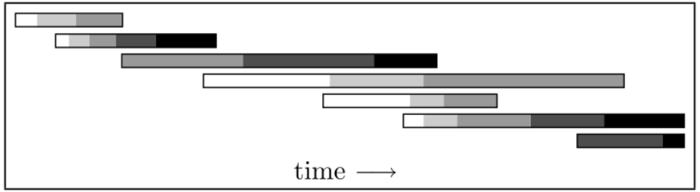
Each runner's fitness level evolves at a different instant in time

Community evolution at learned intervals

Individual user evolution at uniform intervals

A runner's fitness level may evolve at a non-uniform rate

Individual user evolution at learned intervals



Source : J. J. McAuley and J. Leskovec. From amateurs to connoisseurs: modeling the evolution of user expertise through online reviews. In *World Wide Web*, 2013.

Endomondo Data Set

- Free app and website to allow users to keep track of workouts
- Data for each workout
  - Total duration, distance
  - Instantaneous heart rate, duration, distance
  - GPS coordinates
  - Timestamp
- We use only ‘running’ workouts in this work

Activity	Number of Workouts	Number of Users
Running	4413262	1187925
Walking	2154563	558436
Cycling, sport	938108	260828
Cycling, transport	816808	216750
Mountain biking	429906	130603
Step counter	265403	13353
Fitness walking	172440	47775

Baseline Model - Duration

Predicted duration of *i*'th workout of runner *u*

One parameter per runner

Distance of *i*'th workout of runner *u*

$$T'_{ui} = (\alpha + \alpha_u)(\theta_0 + \theta_1 d_{ui})$$

Parameters common to all runners

$$\hat{\Theta} = \arg\min_{\Theta} \frac{1}{|D|} \sum_{T_{ui} \in D} (T'_{ui} - T_{ui})^2 + \lambda \|\Theta\|_2^2$$

Parameters

True duration

Regularizer

Optimize using L-BFGS

Baseline Model - Average Heart-Rate

Predicted Avg. Heart-Rate of *i*'th workout of runner *u*

One parameter per runner

Distance of *i*'th workout of runner *u*

$$H'_{ui} = (\alpha + \alpha_u)(\theta_0 + \theta_1 d_{ui})$$

Parameters common to all runners

$$\hat{\Theta} = \arg\min_{\Theta} \frac{1}{|D|} \sum_{H_{ui} \in D} (H'_{ui} - H_{ui})^2 + \lambda\Omega(\Theta)$$

Parameters

True Avg. HR

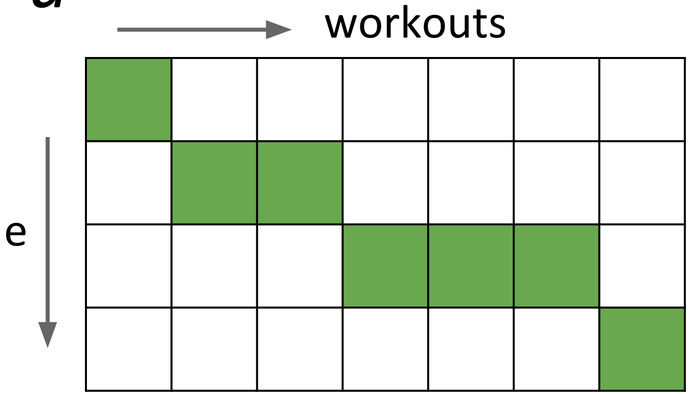
Regularizer

Optimize using L-BFGS

Temporal Evolution *Across* Workouts

- How does a runner's fitness level change from one run to the next?
- Associate an experience level $e_{ui} \in \{1, ..., E\}$ with each workout *i* of runner *u*
- E experience levels
- Monotonicity constraint:

$$\forall u, i, j \qquad t_{ui} \geq t_{uj} \implies e_{ui} \geq e_{uj}$$



Temporal Evolution *Across* Workouts

Model Specification

One parameter per runner
per experience level

$$H'_{ui} = (\alpha_{e_{ui}} + \alpha_{ue_{ui}})(\theta_0 + \theta_1 d_{ui})$$

One parameter per
experience level

Parameters common to all
runners and experience levels

E different models, for E experience levels

Temporal Evolution *Across* Workouts

Training

- Objective

$$(\hat{\Theta}, \hat{\beta}) = \arg\min_{\Theta,\beta} \frac{1}{|D|} \sum_{H_{ui} \in D} (H'_{ui} - H_{ui})^2 + \lambda_1 \Omega_1(\Theta) + \lambda_2 \Omega_2(\Theta)$$

Parameters of the model　　Fitted experience levels　　Regularizers

- Coordinate Descent
    - Optimize parameters Θ given experience levels β : L-BFGS
    - Fit experience levels β given parameters Θ : Dynamic Programming

Temporal Evolution *Across* Workouts

Training

- Regularizers

$$\Omega_1(\Theta) = \sum_{e=1}^{E-1} ||\Theta_e - \Theta_{e+1}||_2^2$$

Penalizes abrupt changes in parameters of successive experience levels

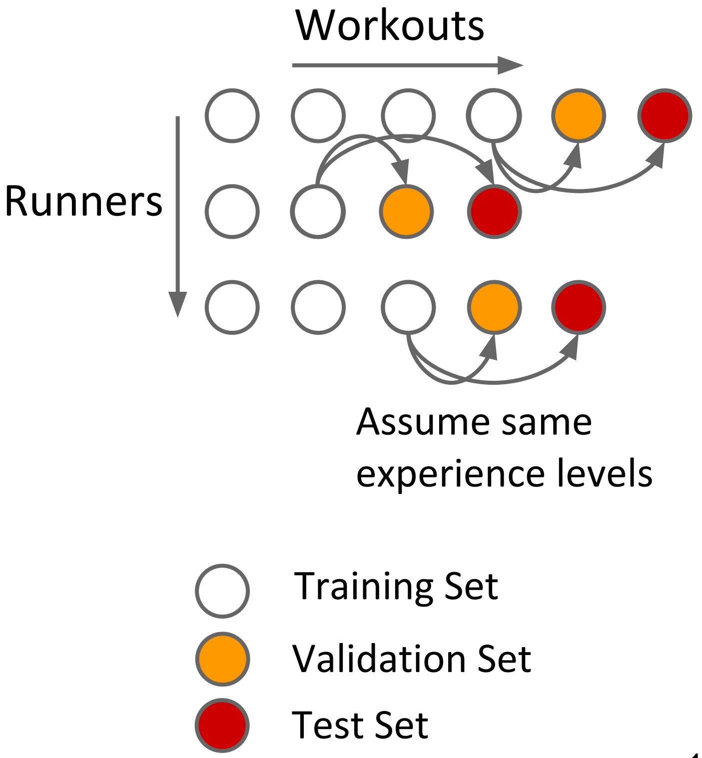
$$\Omega_2(\Theta) = \theta_1^2 + \sum_{e=1}^{E} ||\Theta_e||_2^2$$

Penalizes complexity of the model

Temporal Evolution *Across* Workouts

Selecting validation and test sets

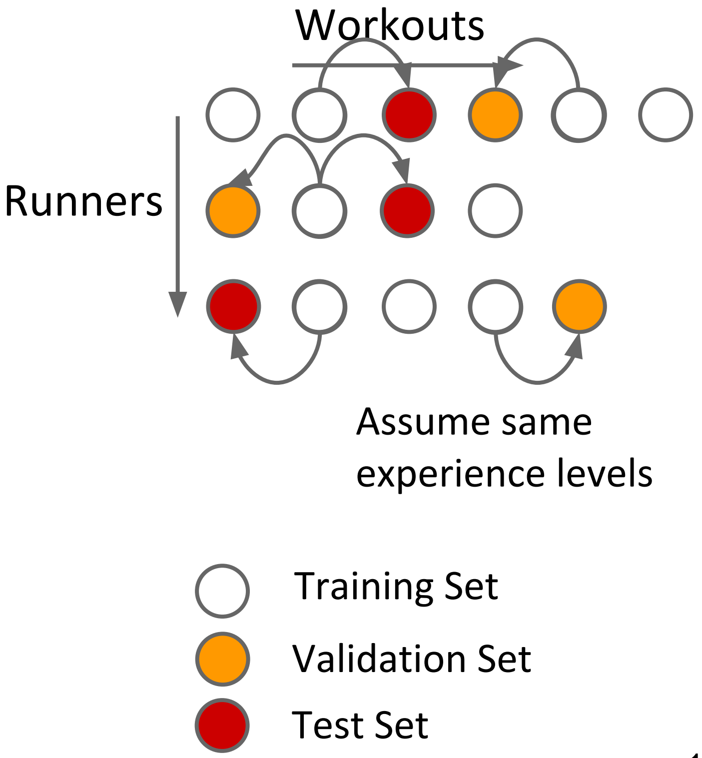
- ‘Final’ mode
  - Last 2 workouts per runner
  - Matches real-life setting
  - Validation and test sets biased towards workouts of experienced runners
  - Assume experience level of last workout for that runner in training set during prediction on validation and test set



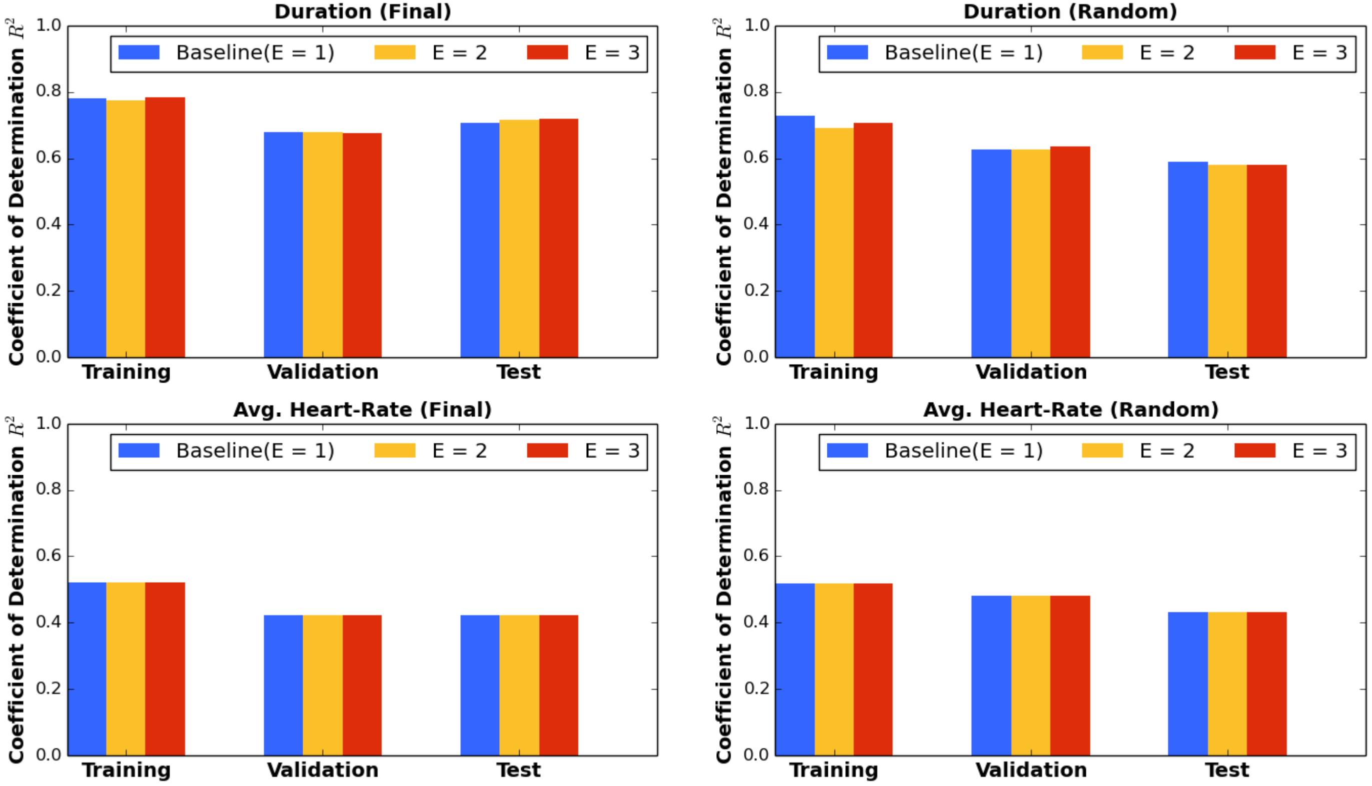
Temporal Evolution *Across* Workouts

Selecting validation and test sets

- ‘Random’ mode
  - 2 randomly selected workouts for each runner
  - Unbiased validation and test sets
  - Assume experience level of workout closest in chronological order in training set during prediction of validation and test set



Results



Set	Examples
Training	743,987
Validation	52,109
Test	52,109

Set	Examples
Training	39,545
Validation	2384
Test	2384

Temporal Evolution *Across* Workouts

Hyper-parameters

Avg. HR	λ_1	λ_2
E = 2 (final)	10000	0.0
E = 3 (final)	100000	0.0
E = 2 (random)	100000	0.0
E = 3 (random)	10000	0.0

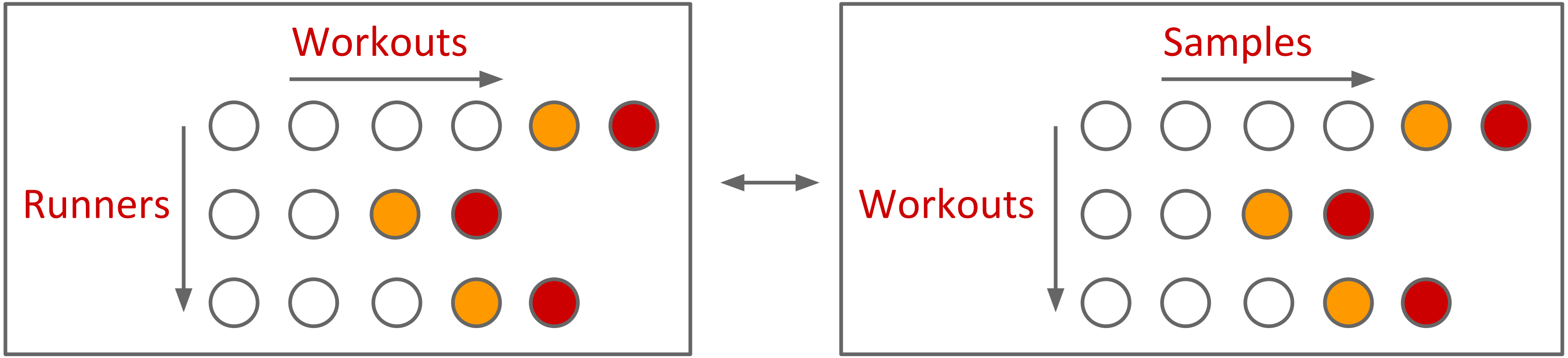
Duration	λ_1	λ_2
E = 2 (final)	10000	0.001
E = 3 (final)	100	0.001
E = 2 (random)	10	0.01
E = 3 (random)	10	0.01

High values of λ_1 mean that the experience levels have very similar parameters indicating that the temporal model reduces to the baseline

Temporal Evolution *Within* Workouts

Apply the same idea on samples within each workout instead of workouts of each runner

- Runners ⟷ Workouts
- Workouts ⟷ Samples
- Experience level ⟷ Tiredness level



Temporal Evolution *Within* Workouts

Apply the same idea on samples within each workout instead of workouts of each runner

- Runners $\longleftrightarrow$ Workouts
- Workouts $\longleftrightarrow$ Samples
- Experience level $\longleftrightarrow$ Tiredness level

Avg. Heart-Rate of *i*'th workout of runner *u*

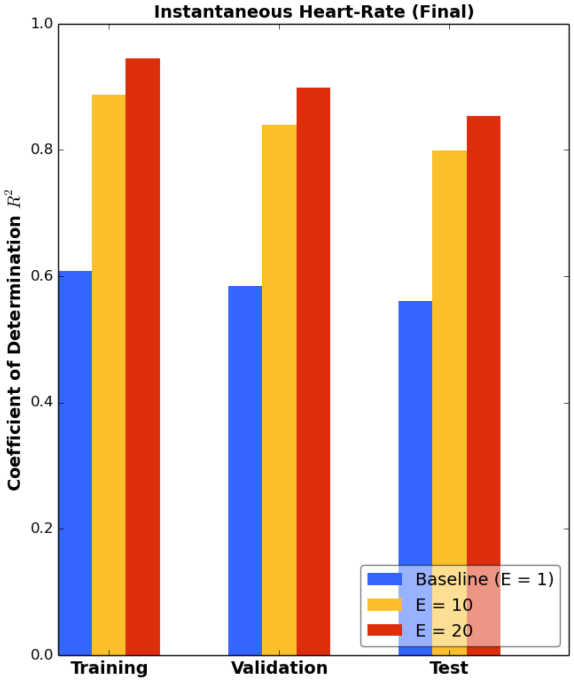
$$H'_{ui} = (\alpha_{e_{ui}} + \alpha_{ue_{ui}})(\theta_0 + \theta_1 d_{ui})$$

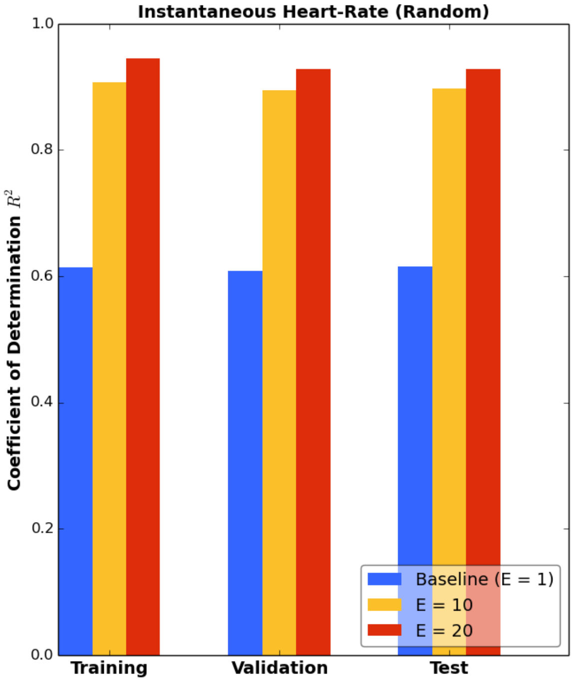
Instantaneous Heart-Rate at *t*'th sample in workout *w*

$$h'_{wt} = (\alpha_{e_{wt}} + \alpha_{we_{wt}})(\theta_0 + \theta_1 d_{wt})$$

Results

Prediction of a single instantaneous heart-rate value





Set	Examples
Training	24,347,765
Validation	83,423
Test	83,423

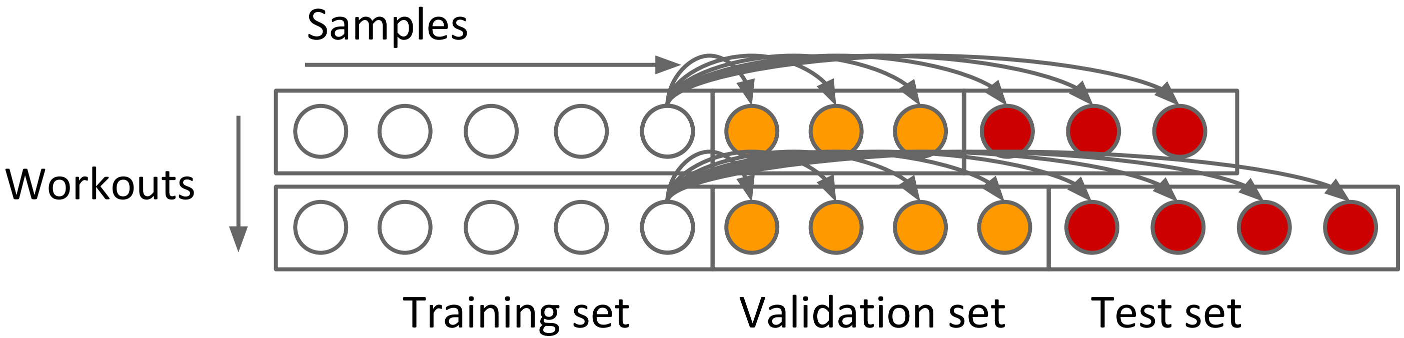
Significant improvement over the baseline

Predicting Several Heart-Rate Values

- How will my heart-rate change in the remaining part of this run?
- Predicting this is useful for runners monitoring their heart-rate during the run

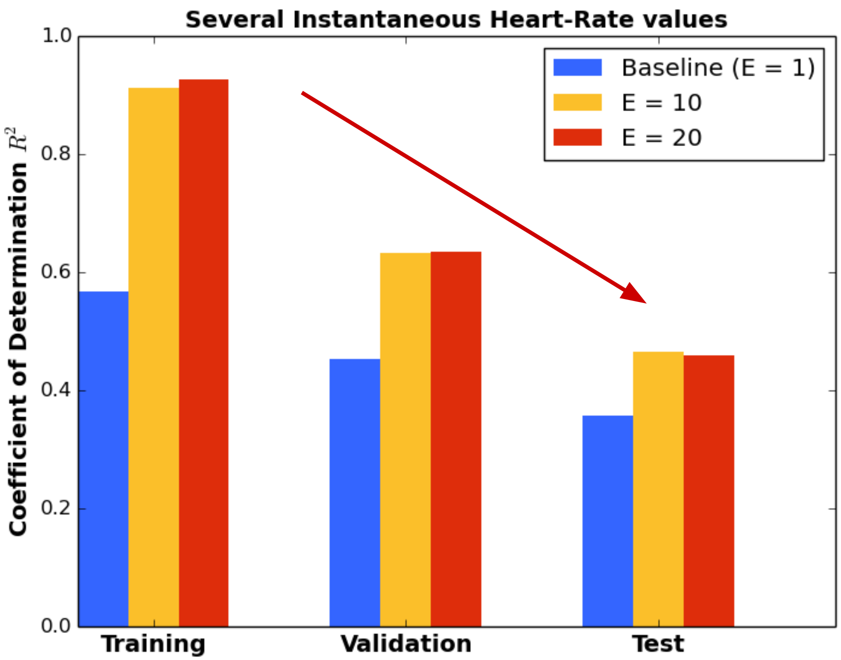
Predicting Several Heart-Rate Values

- How will my heart-rate change in the remaining part of this run?



Learn parameters and fit tiredness levels on training set
Assume last tiredness level in training part for validation and test parts of workout

Predicting Several Heart-Rate Values



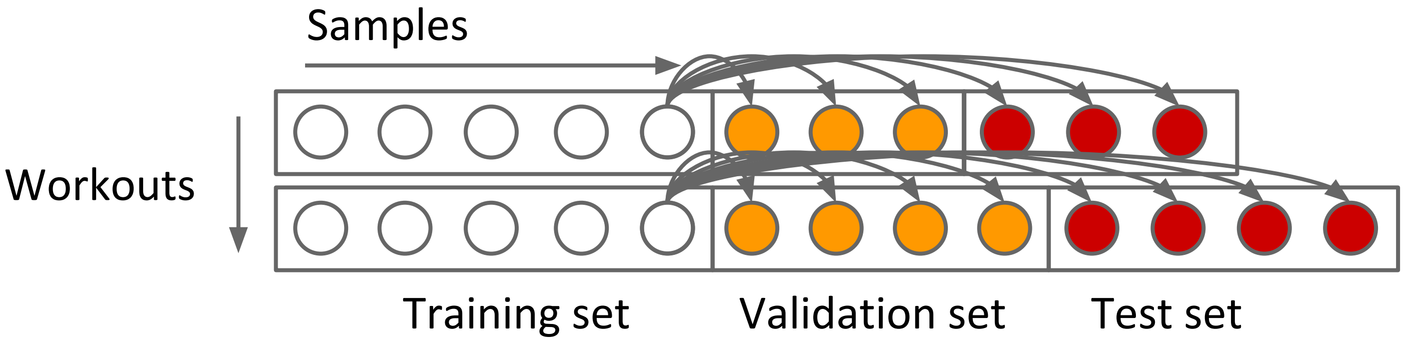
Set	Examples
Training	9,841,317
Validation	7,352,229
Test	7,321,065

Why is training R^2 much higher than validation R^2, which is much higher than test R^2 ?

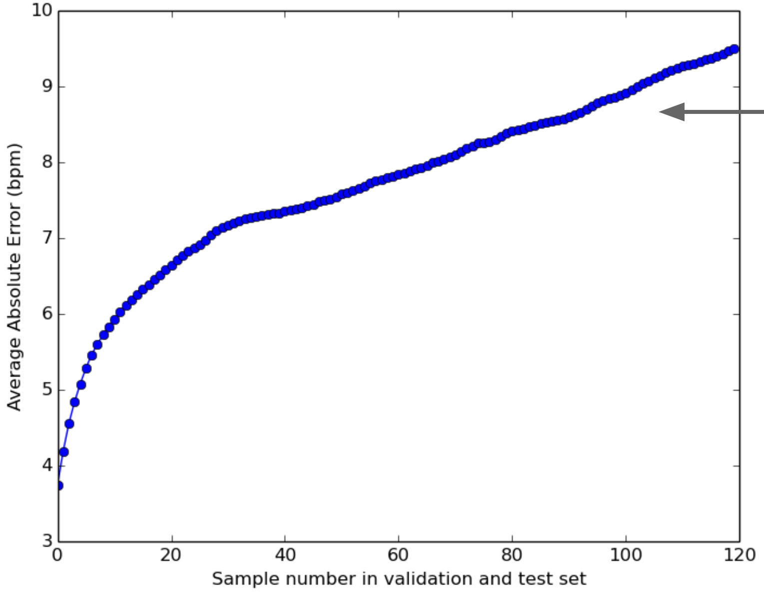
Significant improvement over the baseline

Predicting Several Heart-Rate Values

- We assume that the last tiredness level in the workout fitted during training stays the same for the remaining part of the workout
    - Assumption becomes lesser and lesser likely to hold as we go further in time

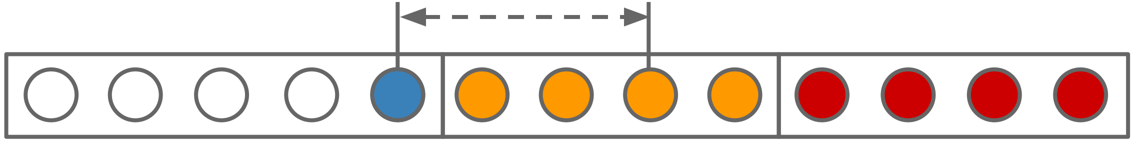


Predicting Several Heart-Rate Values



Error in prediction increases for samples further away from the last sample used for training

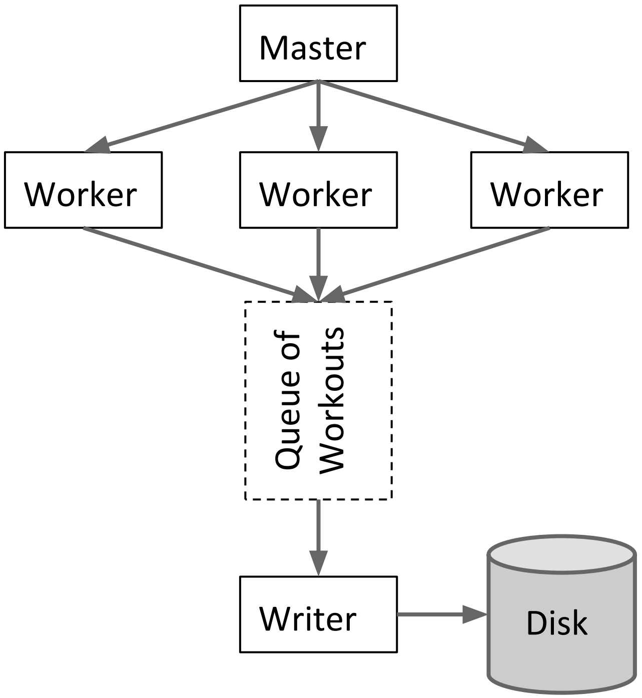
Represents number of samples since the last sample used for training



Training set Validation set Test set

Extracting Relevant Data

- Data available as 200 GB SQL dump after compression
- Over 10 million workouts
- Workout data in HTML and JSON formats
- Parallelization speeds up data extraction



Conclusion

- Significant improvement over baseline when modeling temporal evolution within workouts
- No improvement over baseline when modeling temporal evolution across workouts
  - Not enough data per runner
  - Runners might evolve in ways other than through a change in duration - distance or heart rate - distance relationship

Acknowledgments

I would like to thank my advisor Prof. Julian McAuley for patiently guiding me for the last 3 quarters, throughout the course of my project.

Thank you !

Questions?

Code available on GitHub : https://github.com/yashkarandikar/ucsd-fitness