Can we assess mental health through social media and smart devices?

Addressing bias in methodology and evaluation

A. Tsakalidis, M. Liakata, T. Damoulas, A.I. Cristea

EPSRC
Pioneering research and skills

Introduction

Mental Health Assessment

- Self-reports (time-consuming) → real-time?

Introduction

Mental Health Assessment

- Self-reports (time-consuming) → real-time?

Social Media & Smart Devices

- Real-time sensors of individuals: location, text, moving patterns…
- Relation to mental health?

Introduction

Mental Health Assessment

- Self-reports (time-consuming) → real-time?

Social Media & Smart Devices

- Real-time sensors of individuals: location, text, moving patterns…
- Relation to mental health?

Longitudinal Task Description:

- Small group of subjects monitored over time
- Features: social media, smart devices
- Target: daily self-reported mental health scores

Goal: train models on the features, aiming to predict the current mental health score of an individual

Introduction

Mental Health Assessment

- Self-reports (time-consuming) → real-time?

Social Media & Smart Devices

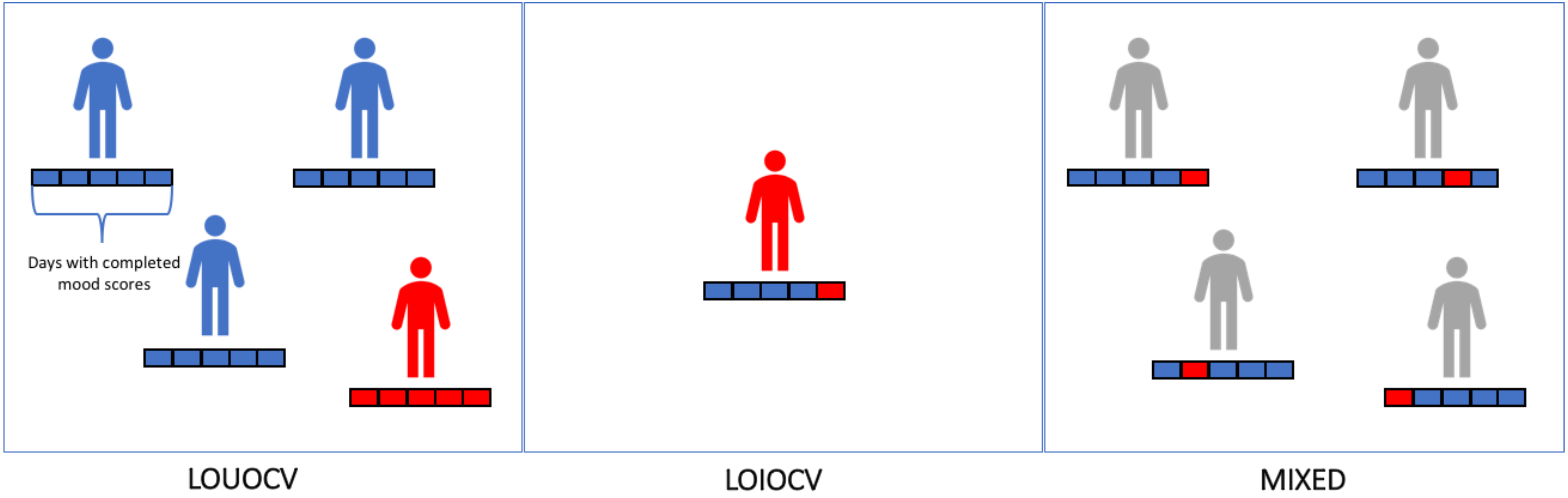
- Real-time sensors of individuals: location, text, moving patterns…
- Relation to mental health?

Longitudinal Task Description:

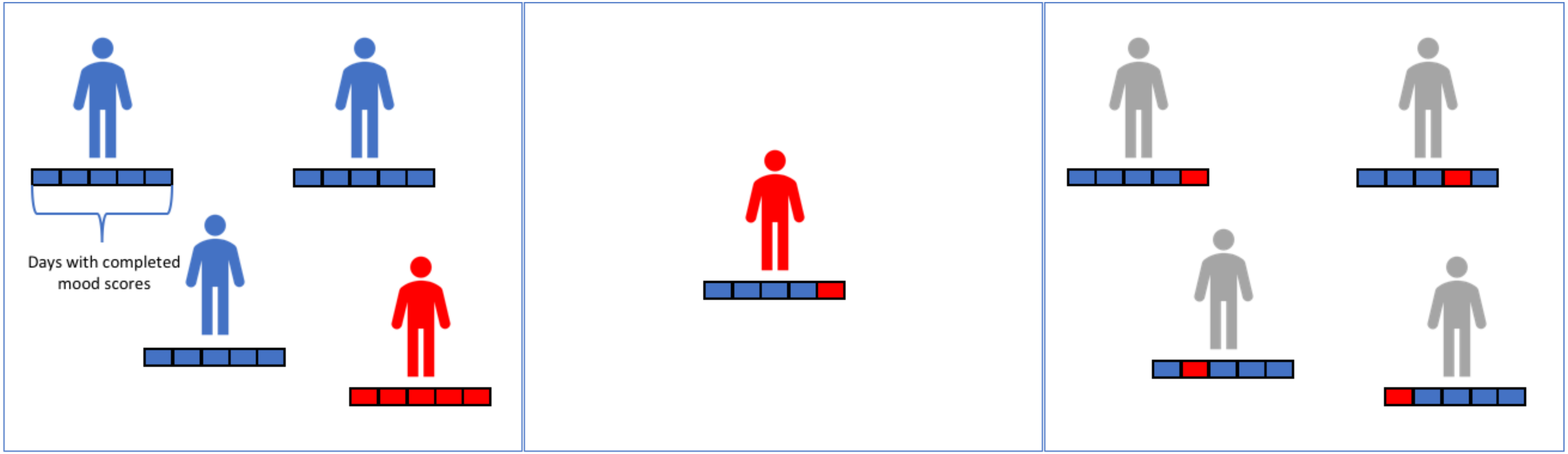
- Small group of subjects monitored over time
- Features: social media, smart devices
- Target: daily self-reported mental health scores
- **Problem: how do we evaluate our models? real-world setting?**

Goal: train models on the features, aiming to predict the current mental health score of an individual

Types of Evaluation



Types of Evaluation

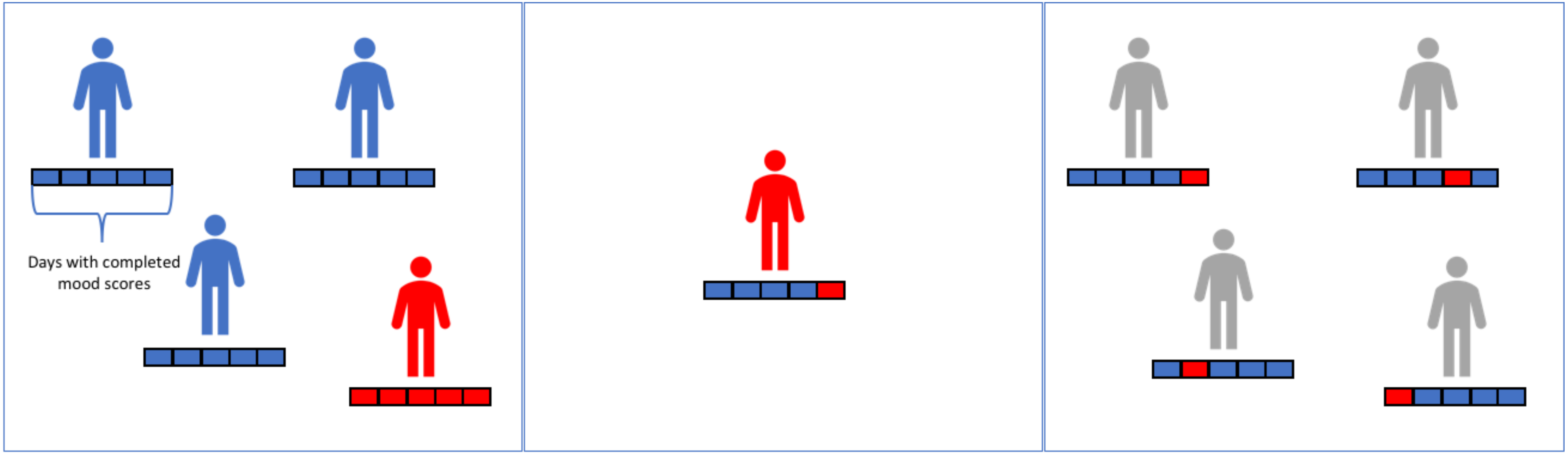


LOUOCV LOIOCV MIXED

Goal: Generalise to new users

Problem: few users…

Types of Evaluation



LOUOCV LOIOCV MIXED

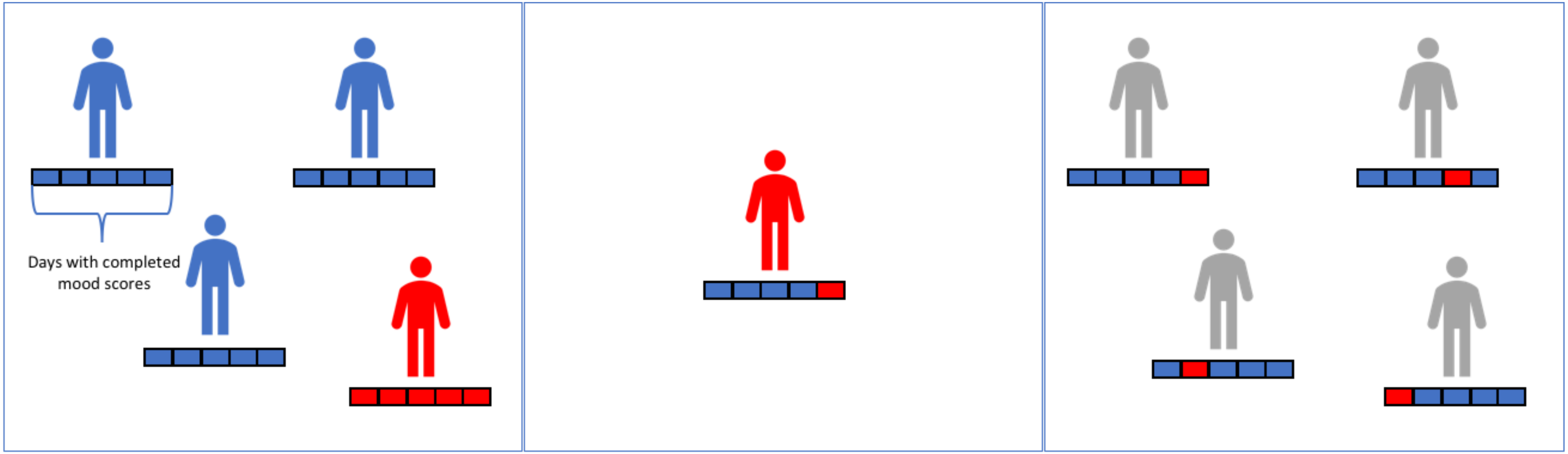
Goal: Generalise to new users

Problem: few users…

Goal: Personalised models

Problem: few instances…

Types of Evaluation



LOUOCV LOIOCV MIXED

LOUOCV

Goal: Generalise to new users

Problem: few users…

LOIOCV

Goal: Personalised models

Problem: few instances…

MIXED

Goal: Generalise to certain users only

Problem: identify the user & infer his/her ”average” score

Problem Statement

Problem Statement

P1: Training on past values of the target [1]

- Using past days' target scores as features
- Problems:
  - LOUOCV: cannot assess mental health of a new user
  - LOIOCV: target score in test instance is used as a feature in another training example!

Problem Statement

P1: Training on past values of the target [1]

- Using past days' target scores as features
- Problems:
  - LOUOCV: cannot assess mental health of a new user
  - LOIOCV: target score in test instance is used as a feature in another training example!

P2: Inferring test set labels [1, 2]

- Problem in LOIOCV:
  - Creating overlapping instances (e.g., total walking distance over the past 3 days)
  - Test set instance features: correlated with the (temporally) close instances in the train set
  - What if the target is also (temporally) correlated?

Problem Statement

P1: Training on past values of the target [1]

- Using past days' target scores as features
- Problems:
  - LOUOCV: cannot assess mental health of a new user
  - LOIOCV: target score in test instance is used as a feature in another training example!

P2: Inferring test set labels [1, 2]

- Problem in LOIOCV:
  - Creating overlapping instances (e.g., total walking distance over the past 3 days)
  - Test set instance features: correlated with the (temporally) close instances in the train set
  - What if the target is also (temporally) correlated?

P3: Predicting users instead of mood scores [3,4]

- Problem in MIXED:
  - instances of the same user in train/test set => identify the user in the test set?

Problem Statement

- **P1: Training on past values of the target**
- **P2: Inferring test set labels**
- **P3: Predicting users instead of mood scores**

- **Our goal:**
  - Follow past SOTA for each of the identified problems (P1, P2, P3)
    - Pre-processing, model building, feature selection…
  - Test them in different datasets *under a real-world setting*
  - Demonstrate the issues through experimentation
  - Propose directions for future work

Datasets and Features

- **Dataset 1**
  - 27 subjects
  - ~4 months of data
  - **3 targets** (positive, negative [10-50], wellbeing [14-70]) [6, 7]
  - **textual features** (posts & private messages from social media & SMS):
    ngrams, lexicons, word clusters, word embeddings, count-based (e.g., number of SMS)

- **Dataset 2** [5]
  - 44 subjects
  - 10 weeks of data
  - **1 target** (stress [0-4])
  - **smartphone features**:
    % of samples for different activities and audio modes;
    number/duration of: conversations, phone in dark environment, phone locked, phone charging

P1: Training on past values of the target

LiKamWa et al. [1] used the previous two past target scores as features

- LOUOCV: demands input by the new user
- LOIOCV: target in test instance used as feature in training set!

Feature extraction performed over past 3 days (overlapping instances)

P1: Training on past values of the target

LiKamWa et al. [1] used the previous two past target scores as features

- LOUOCV: demands input by the new user
- LOIOCV: target in test instance used as feature in training set!

Feature extraction performed over past 3 days (overlapping instances)

Evaluation

LOUOCV, LOIOCV

AVG: always predicting user average

LAST: last entered mood score

-feat: model trained on the past two target scores only

-mood: model trained on sensor data only

Results: P1 (train on target values)

	positive		negative		wellbeing		stress		LiKamWa et al. [1]	
	MSE	**accuracy**	**MSE**	**accuracy**	**MSE**	**accuracy**	**MSE**	**accuracy**	**MSE**	**accuracy**
LOIOCV	15.96	84.5	11.64	87.1	20.94	89.0	1.07	47.3	0.08	93.0
LOUOCV	36.77	63.4	31.99	68.3	51.08	72.8	0.81	45.4	0.29	66.5
AVG	29.89	71.8	27.80	73.1	41.14	78.9	0.70	51.6	0.24	73.5
LAST	43.44	60.4	38.22	63.2	55.73	71.6	1.15	51.5	0.34	63.0
-feat	33.40	67.2	28.60	72.3	45.66	76.6	0.81	49.8	0.27	70.5
-mood	113.30	30.9	75.27	44.5	138.67	42.5	1.08	44.4	N/A	N/A

Results: P1 (train on target values)

	positive		negative		wellbeing		stress		LiKamWa et al. [1]	
	MSE	**accuracy**	**MSE**	**accuracy**	**MSE**	**accuracy**	**MSE**	**accuracy**	**MSE**	**accuracy**
LOIOCV	15.96	84.5	11.64	87.1	20.94	89.0	1.07	47.3	0.08	93.0
LOUOCV	36.77	63.4	31.99	68.3	51.08	72.8	0.81	45.4	0.29	66.5
AVG	29.89	71.8	27.80	73.1	41.14	78.9	0.70	51.6	0.24	73.5
LAST	43.44	60.4	38.22	63.2	55.73	71.6	1.15	51.5	0.34	63.0
-feat	33.40	67.2	28.60	72.3	45.66	76.6	0.81	49.8	0.27	70.5
-mood	113.30	30.9	75.27	44.5	138.67	42.5	1.08	44.4	N/A	N/A

LOUOCV

- Always predicting the average mood is better (AVG); results much worse if no target scores are used (-mood).

Results: P1 (train on target values)

	positive		negative		wellbeing		stress		LiKamWa et al. [1]	
	MSE	accuracy	MSE	accuracy	MSE	accuracy	MSE	accuracy	MSE	accuracy
LOIOCV	15.96	84.5	11.64	87.1	20.94	89.0	1.07	47.3	0.08	93.0
LOUOCV	36.77	63.4	31.99	68.3	51.08	72.8	0.81	45.4	0.29	66.5
AVG	29.89	71.8	27.80	73.1	41.14	78.9	0.70	51.6	0.24	73.5
LAST	43.44	60.4	38.22	63.2	55.73	71.6	1.15	51.5	0.34	63.0
-feat	33.40	67.2	28.60	72.3	45.66	76.6	0.81	49.8	0.27	70.5
-mood	113.30	30.9	75.27	44.5	138.67	42.5	1.08	44.4	N/A	N/A

LOUOCV

- Always predicting the average mood is better (AVG); results much worse if no target scores are used (-mood).

LOIOCV

- Performance much better! But instances are created in overlapping time windows (3-days).
  What if our target is also correlated with respect to time?

P2: Inferring test set labels

Instance generation: Canzian & Musolesi [2] extracted features from overlapping time windows

- T_{HIST} = {1, …, 14} days before the completion of a mood form
- For high T_{HIST}, the features are highly correlated!

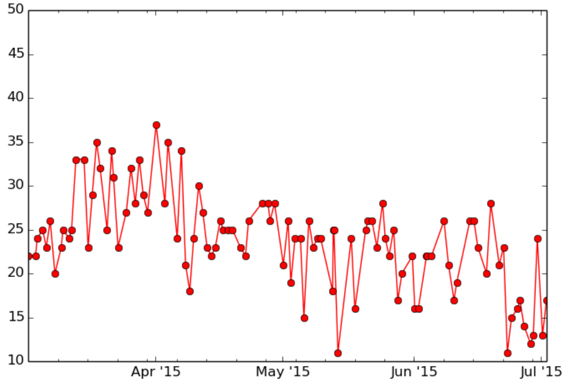
P2: Inferring test set labels

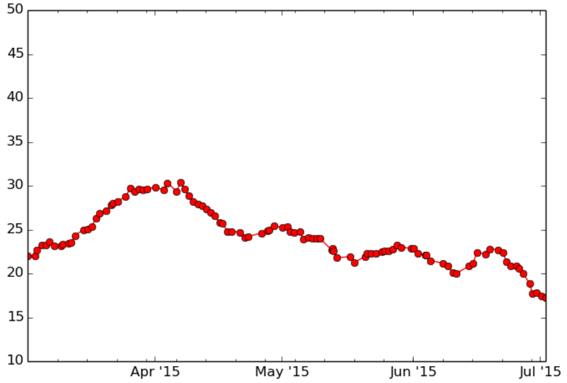
Instance generation: Canzian & Musolesi [2] extracted features from overlapping time windows

- T_{HIST} = {1, …, 14} days before the completion of a mood form
- For high T_{HIST}, the features are highly correlated!

Target pre-processing: moving averages

- Target is now also temporally correlated





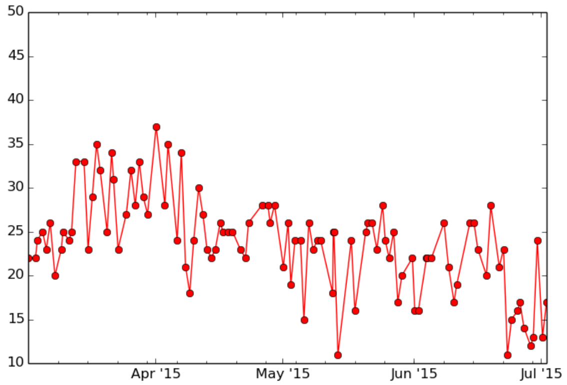
P2: Inferring test set labels

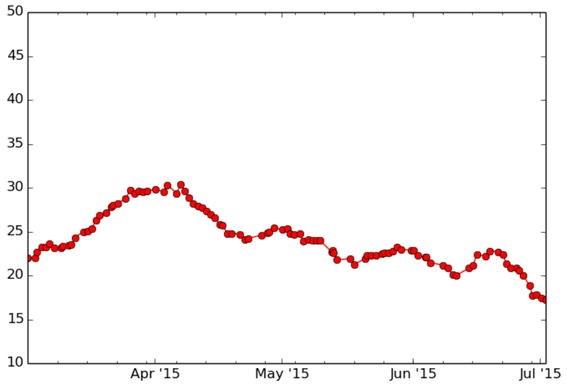
Instance generation: Canzian & Musolesi [2] extracted features from overlapping time windows

- T_{HIST} = {1, …, 14} days before the completion of a mood form
- For high T_{HIST}, the features are highly correlated!

Target pre-processing: moving averages

- Target is now also temporally correlated





Evaluation

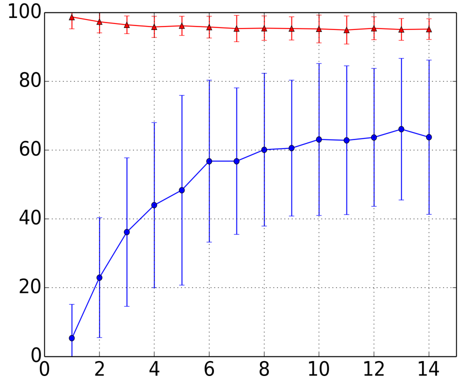
- LOIOCV
- binary (high/low) classification
- wider T_{HIST} => better results

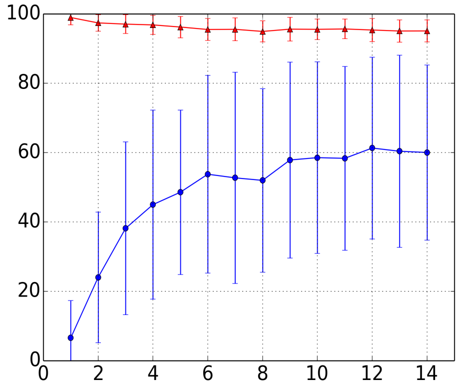
Results: P2 (inferring test set labels)

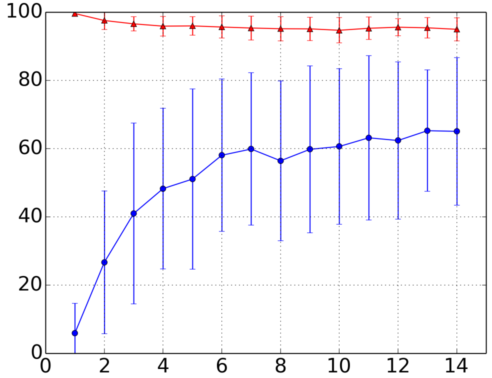
LOIOCV

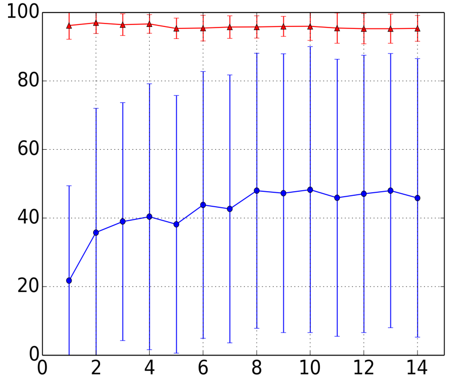
Same findings with [2]:

- *Sensitivity* increases with larger window size
- *Specificity* remains stable at high values





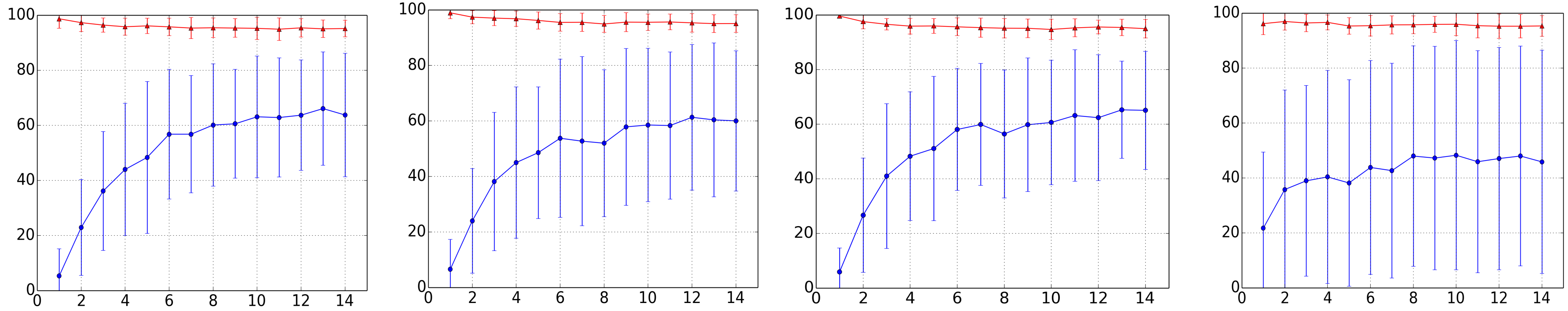




Results: P2 (inferring test set labels)

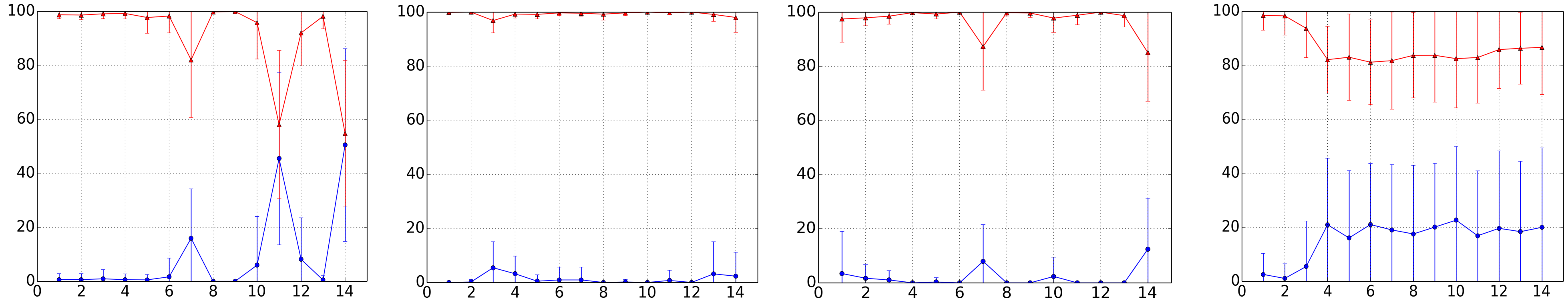
LOIOCV

Same findings with [2]:
- *Sensitivity* increases with larger window size
- *Specificity* remains stable at high values



LOUOCV

- Window size does not affect *sensitivity*
- Increase in sensitivity is accompanied by sharp drops in *specificity*

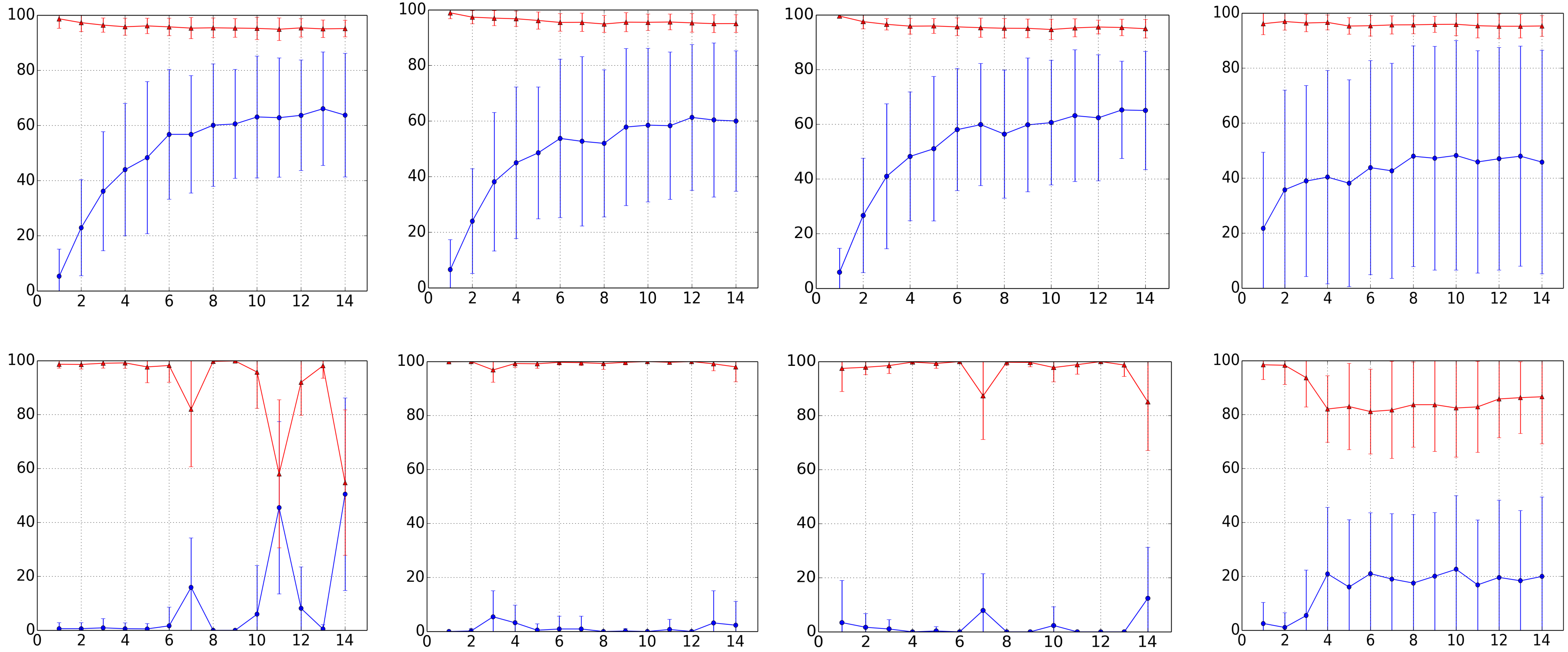


Results: P2 (inferring test set labels)

LOIOCV

Same findings with [2]:
- *Sensitivity* increases with larger window size
- *Specificity* remains stable at high values

LOUOCV
- Window size does not affect *sensitivity*
- Increase in sensitivity is accompanied by sharp drops in *specificity*



LOIOCV

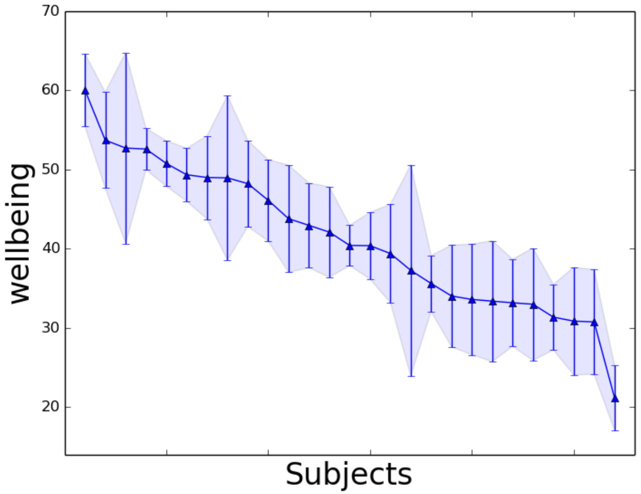
Comparison of model (FEAT) with T_{HIST} = 14 days against naïve baselines

FEAT:	64.02	60.03	65.06	45.86
DATE:	59.68	62.75	63.29	46.99
LAST:	67.37	69.08	66.05	58.20
RAND:	64.22	60.88	64.87	45.79

P3: Predicting users instead of mood scores

Tsakalidis et al. [4] evaluated regression models under MIXED

- Per-user (textual) feature normalisation => better performance
- LOUOCV/LOIOCV?



Jaques et al. [3] separated instances based on high/low scores across all subjects (binary classification)

- Separating high/low on a per-user basis?
- LOUOCV/LOIOCV?

Results: P3 (predicting the user)

Experiment 1 (regression) [4]

	positive		negative		wellbeing		stress	
	R^2	ε	R^2	ε	R^2	ε	R^2	ε
MIXED_{+}								
MIXED_{-}								
LOIOCV_{+}								
LOIOCV_{-}								
LOUOCV_{+}								
LOUOCV_{-}								

Results: P3 (predicting the user)

	positive		negative		wellbeing		stress	
	R^2	ε	R^2	ε	R^2	ε	R^2	ε
MIXED_+	0.43	6.91	0.25	6.49	0.48	8.04	0.02	1.03
MIXED_-	0.13	8.50	0.00	7.52	0.31	10.33	0.03	1.03
LOIOCV_+								
LOIOCV_-								
LOUOCV_+								
LOUOCV_-								

Experiment 1 (regression) [4]

MIXED: Effect of (per-user) feature normalisation

Results: P3 (predicting the user)

	positive		negative		wellbeing		stress	
	R^2	ε	R^2	ε	R^2	ε	R^2	ε
$MIXED_+$	0.43	6.91	0.25	6.49	0.48	8.04	0.02	1.03
$MIXED_-$	0.13	8.50	0.00	7.52	0.31	10.33	0.03	1.03
$LOIOCV_+$	-0.03	5.20	-0.04	5.05	-0.03	6.03	-0.08	0.91
$LOIOCV_-$	-0.03	5.20	-0.04	5.05	-0.03	6.03	-0.08	0.91
$LOUOCV_+$								
$LOUOCV_-$								

Experiment 1 (regression) [4]

MIXED: Effect of (per-user) feature normalisation

LOIOCV: Worse than the average mood predictor

Results: P3 (predicting the user)

	positive		negative		wellbeing		stress	
	R^2	ε	R^2	ε	R^2	ε	R^2	ε
$MIXED_+$	0.43	6.91	0.25	6.49	0.48	8.04	0.02	1.03
$MIXED_-$	0.13	8.50	0.00	7.52	0.31	10.33	0.03	1.03
$LOIOCV_+$	-0.03	5.20	-0.04	5.05	-0.03	6.03	-0.08	0.91
$LOIOCV_-$	-0.03	5.20	-0.04	5.05	-0.03	6.03	-0.08	0.91
$LOUOCV_+$	-4.19	8.98	-1.09	7.24	-4.66	10.61	-0.67	1.01
$LOUOCV_-$	-4.38	8.98	-1.41	7.23	-4.62	10.62	-0.69	1.02

Experiment 1 (regression) [4]

MIXED: Effect of (per-user) feature normalisation

LOIOCV: Worse than the average mood predictor

LOUOCV: Results rather poor

Results: P3 (predicting the user)

Experiment 1 (regression) [4]

	positive		negative		wellbeing		stress	
	R^2	ε	R^2	ε	R^2	ε	R^2	ε
$MIXED_+$	0.43	6.91	0.25	6.49	0.48	8.04	0.02	1.03
$MIXED_-$	0.13	8.50	0.00	7.52	0.31	10.33	0.03	1.03
$LOIOCV_+$	-0.03	5.20	-0.04	5.05	-0.03	6.03	-0.08	0.91
$LOIOCV_-$	-0.03	5.20	-0.04	5.05	-0.03	6.03	-0.08	0.91
$LOUOCV_+$	-4.19	8.98	-1.09	7.24	-4.66	10.61	-0.67	1.01
$LOUOCV_-$	-4.38	8.98	-1.41	7.23	-4.62	10.62	-0.69	1.02

MIXED: Effect of (per-user) feature normalisation

LOIOCV: Worse than the average mood predictor

LOUOCV: Results rather poor

Experiment 2 (classification) [3]

	positive		negative		wellbeing		stress	
	UNIQ	PERS	UNIQ	PERS	UNIQ	PERS	UNIQ	PERS
MIXED	65.69	51.54	60.68	55.79	68.14	51.00	61.75	56.44
LOIOCV	78.22	51.79	84.86	53.63	88.06	52.89	73.54	55.35
LOUOCV	47.36	50.74	42.41	52.45	45.57	50.10	49.77	55.11

UNIQ: Labelling instances based on high/low mood scores across all users [3]

PERS: Labelling instances on a per-user basis

Conclusions similar to E1

Results: P3 (predicting the user)

	positive		negative		wellbeing		stress	
	R^2	ε	R^2	ε	R^2	ε	R^2	ε
$MIXED_+$	0.43	6.91	0.25	6.49	0.48	8.04	0.02	1.03
$MIXED_-$	0.13	8.50	0.00	7.52	0.31	10.33	0.03	1.03
$LOIOCV_+$	-0.03	5.20	-0.04	5.05	-0.03	6.03	-0.08	0.91
$LOIOCV_-$	-0.03	5.20	-0.04	5.05	-0.03	6.03	-0.08	0.91
$LOUOCV_+$	-4.19	8.98	-1.09	7.24	-4.66	10.61	-0.67	1.01
$LOUOCV_-$	-4.38	8.98	-1.41	7.23	-4.62	10.62	-0.69	1.02

Experiment 1 (regression) [4]

MIXED: Effect of (per-user) feature normalisation

LOIOCV: Worse than the average mood predictor

LOUOCV: Results rather poor

	positive		negative		wellbeing		stress	
	UNIQ	PERS	UNIQ	PERS	UNIQ	PERS	UNIQ	PERS
MIXED	65.69	51.54	60.68	55.79	68.14	51.00	61.75	56.44
LOIOCV	78.22	51.79	84.86	53.63	88.06	52.89	73.54	55.35
LOUOCV	47.36	50.74	42.41	52.45	45.57	50.10	49.77	55.11

Experiment 2 (classification) [3]

UNIQ: Labelling instances based on high/low mood scores across all users [3]

PERS: Labelling instances on a per-user basis

Conclusions similar to E1

Results: P3 (predicting the user)

	positive		negative		wellbeing		stress	
	R^2	ε	R^2	ε	R^2	ε	R^2	ε
$MIXED_+$	0.43	6.91	0.25	6.49	0.48	8.04	0.02	1.03
$MIXED_-$	0.13	8.50	0.00	7.52	0.31	10.33	0.03	1.03
$LOIOCV_+$	-0.03	5.20	-0.04	5.05	-0.03	6.03	-0.08	0.91
$LOIOCV_-$	-0.03	5.20	-0.04	5.05	-0.03	6.03	-0.08	0.91
$LOUOCV_+$	-4.19	8.98	-1.09	7.24	-4.66	10.61	-0.67	1.01
$LOUOCV_-$	-4.38	8.98	-1.41	7.23	-4.62	10.62	-0.69	1.02

Experiment 1 (regression) [4]

MIXED: Effect of (per-user) feature normalisation

LOIOCV: Worse than the average mood predictor

LOUOCV: Results rather poor

	positive		negative		wellbeing		stress	
	UNIQ	**PERS**	**UNIQ**	**PERS**	**UNIQ**	**PERS**	**UNIQ**	**PERS**
MIXED	65.69	51.54	60.68	55.79	68.14	51.00	61.75	56.44
LOIOCV	78.22	51.79	84.86	53.63	88.06	52.89	73.54	55.35
LOUOCV	47.36	50.74	42.41	52.45	45.57	50.10	49.77	55.11

Experiment 2 (classification) [3]

UNIQ: Labelling instances based on high/low mood scores across all users [3]

PERS: Labelling instances on a per-user basis

Conclusions similar to E1

Conclusion

- Assessing mental health through smart devices & social media: hard!
- Current SOTA does not follow a real-world setting
  - Important for practitioners!
  - Conclusions reached might be wrong!

Conclusion

- Assessing mental health through smart devices & social media: hard!
- Current SOTA does not follow a real-world setting!
  - Important for practitioners!
  - Conclusions reached might be wrong!

Proposal for Future Directions

- Types of evaluation: {LOUOCV, LOIOCV}
- Demographic information
- Transfer learning
  - Few users with different behaviour
- Latent Feature Representations

Thank you!

Any questions?

Acknowledgements: Supported by the EPSRC through the University of Warwick's CDT in Urban Science and Progress (EP/L016400/1) and through The Alan Turing Institute (EP/N510129/1).

We would like to thank the anonymous reviewers for their detailed feedback and the authors of the works that were analysed in our paper for the fruitful discussions over several aspects of the presented challenges.

References

[1] LiKamWa, R., Liu, Y., Lane, N.D. and Zhong, L., 2013, June. Moodscope: Building a mood sensor from smartphone usage patterns. In *Proceeding of the 11th annual international conference on Mobile systems, applications, and services* (pp. 389-402). ACM.

[2] Canzian, L. and Musolesi, M., 2015, September. Trajectories of depression: unobtrusive monitoring of depressive states by means of smartphone mobility traces analysis. In *Proceedings of the 2015 ACM international joint conference on pervasive and ubiquitous computing* (pp. 1293-1304). ACM.

[3] Jaques, N., Taylor, S., Azaria, A., Ghandeharioun, A., Sano, A. and Picard, R., 2015, September. Predicting students' happiness from physiology, phone, mobility, and behavioral data. In *Affective computing and intelligent interaction (ACII), 2015 international conference on* (pp. 222-228). IEEE.

[4] Tsakalidis, A., Liakata, M., Damoulas, T., Jellinek, B., Guo, W. and Cristea, A., 2016. Combining heterogeneous user generated data to sense well-being. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers* (pp. 3007-3018).

[5] Wang, R., Chen, F., Chen, Z., Li, T., Harari, G., Tignor, S., Zhou, X., Ben-Zeev, D. and Campbell, A.T., 2014, September. StudentLife: assessing mental health, academic performance and behavioral trends of college students using smartphones. In *Proceedings of the 2014 ACM international joint conference on pervasive and ubiquitous computing* (pp. 3-14). ACM.

[6] Watson, D., Clark, L.A. and Tellegen, A., 1988. Development and validation of brief measures of positive and negative affect: the PANAS scales. *Journal of personality and social psychology*, *54*(6), p.1063.

[7] Tennant, R., Hiller, L., Fishwick, R., Platt, S., Joseph, S., Weich, S., Parkinson, J., Secker, J. and Stewart-Brown, S., 2007. The Warwick-Edinburgh mental well-being scale (WEMWBS): development and UK validation. *Health and Quality of life Outcomes*, *5*(1), p.63.