The Alan Turing Institute

WARWICK
THE UNIVERSITY OF WARWICK

Assessing Mental Health through Social Media and Smart Devices

Current State, Challenges and Future Directions

EPSRC
Pioneering research and skills

Adam Tsakalidis

Overview

- **Problem Definition**
- **Part 1:** Assessing Mental Health with Smart Devices and Social Media
  - Dataset Collection
  - Experiments and Results
  - Discussion
- **Part 2:** Addressing Bias in Methodology and Evaluation
  - Real-world Challenges
  - Experiments and Results
- **Conclusion & Future Directions**

Overview

- **Problem Definition**
- **Part 1:** Assessing Mental Health with Smart Devices and Social Media
  - Dataset Collection
  - Experiments and Results
  - Discussion
- **Part 2:** Addressing Bias in Methodology and Evaluation
  - Real-world Challenges
  - Experiments and Results
- **Conclusion & Future Directions**

Introduction

Mental Health

- *"The foundation for well-being and effective functioning for an individual and for a community"* [Herrman et al., 2005]
- Low motivation, lack of satisfaction, low productivity…

Assessment

- Face-to-face assessment
- Self-reports (time-consuming) → real-time?

Social Media & Smart Devices

- Real-time sensors of individuals
- Related to mental health?

Background

Static

Dynamic

Background

Static

- Goal: find users with poor mental health
- User classification
  - e.g., PTSD vs Control Group

✓ Large-scale

✓ "Cheap"

▪ Well-defined classes

▪ Non-longitudinal

Dynamic

Background

Static

- Goal: find users with poor mental health
- User classification
  - e.g., PTSD vs Control Group

✓ Large-scale

✓ "Cheap"

▪ Well-defined classes

▪ Non-longitudinal

Dynamic

- Goal: real-time assessment
- Same users over time
  - e.g., my mood on Monday, Tuesday, …

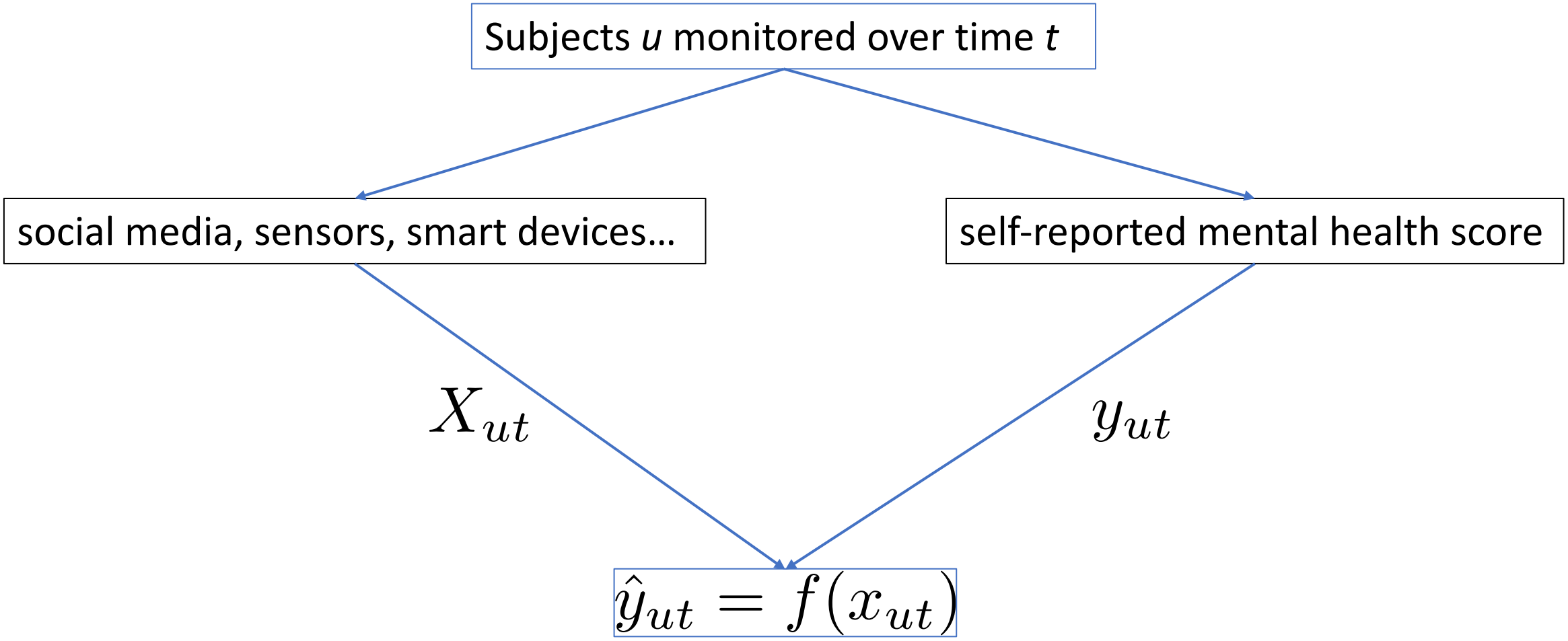
✓ Longitudinal

✓ Real-time monitoring

▪ Small-scale

▪ "Expensive"

Task Definition

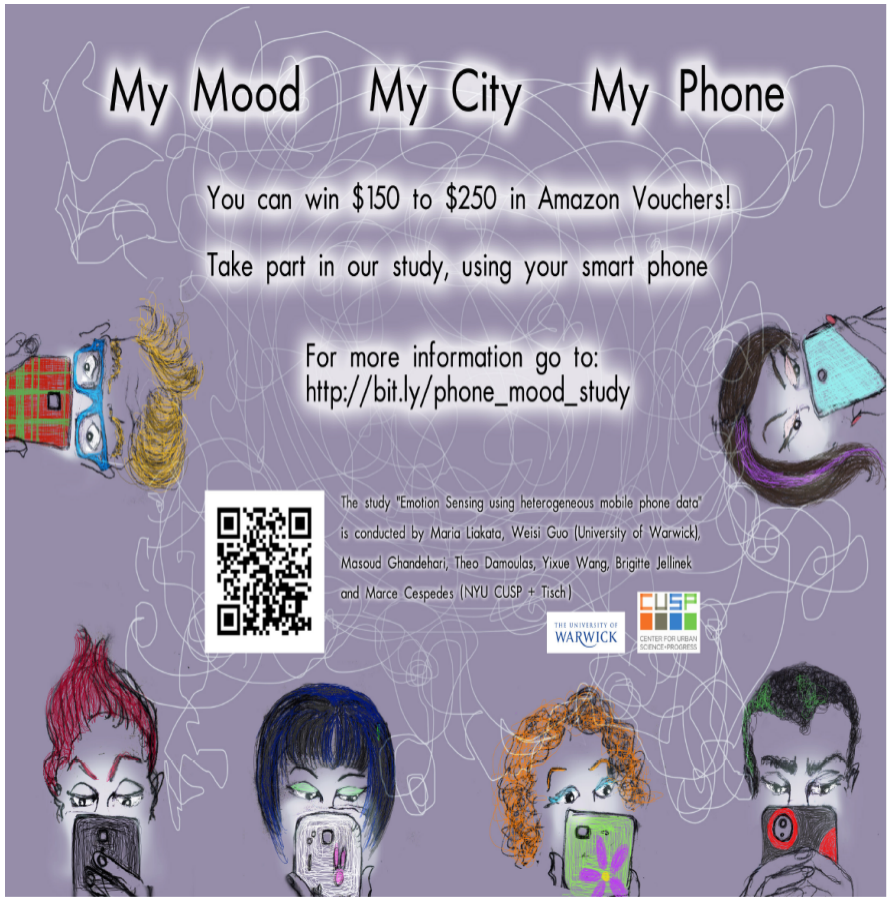


Overview

- **Problem Definition**
- **Part 1:** Assessing Mental Health with Smart Devices and Social Media
  - Dataset Collection
  - Experiments and Results
  - Discussion
- **Part 2:** Addressing Bias in Methodology and Evaluation
  - Real-world Challenges
  - Experiments and Results
  - Future Directions
- **Conclusion Future Directions**

Dataset

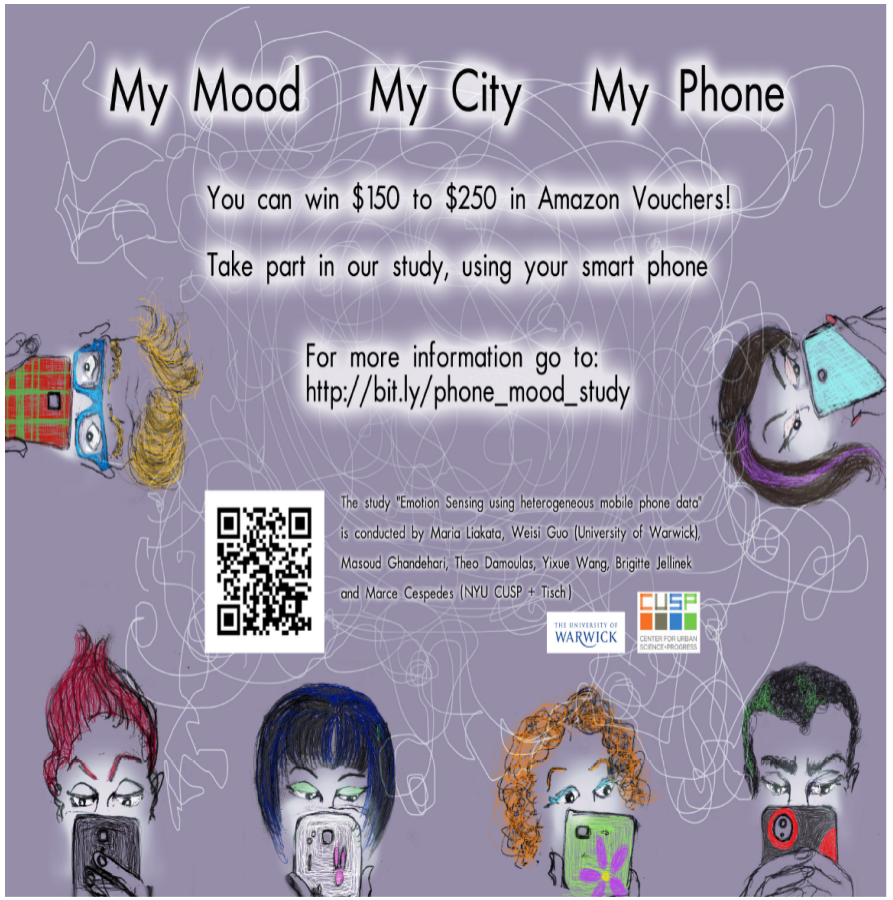
29 users, ~4 months



Dataset

29 users, ~4 months

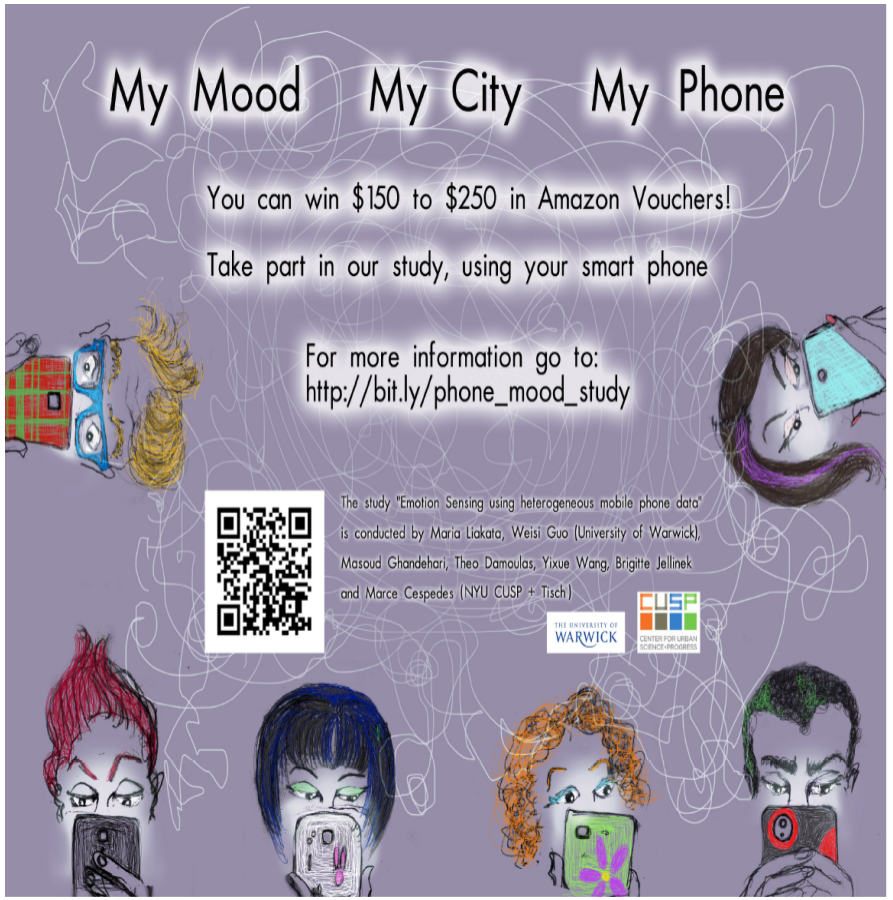
- TEXT: posts/messages (>110K)
  - Facebook, Twitter, Google+
  - SMS



Dataset

29 users, ~4 months

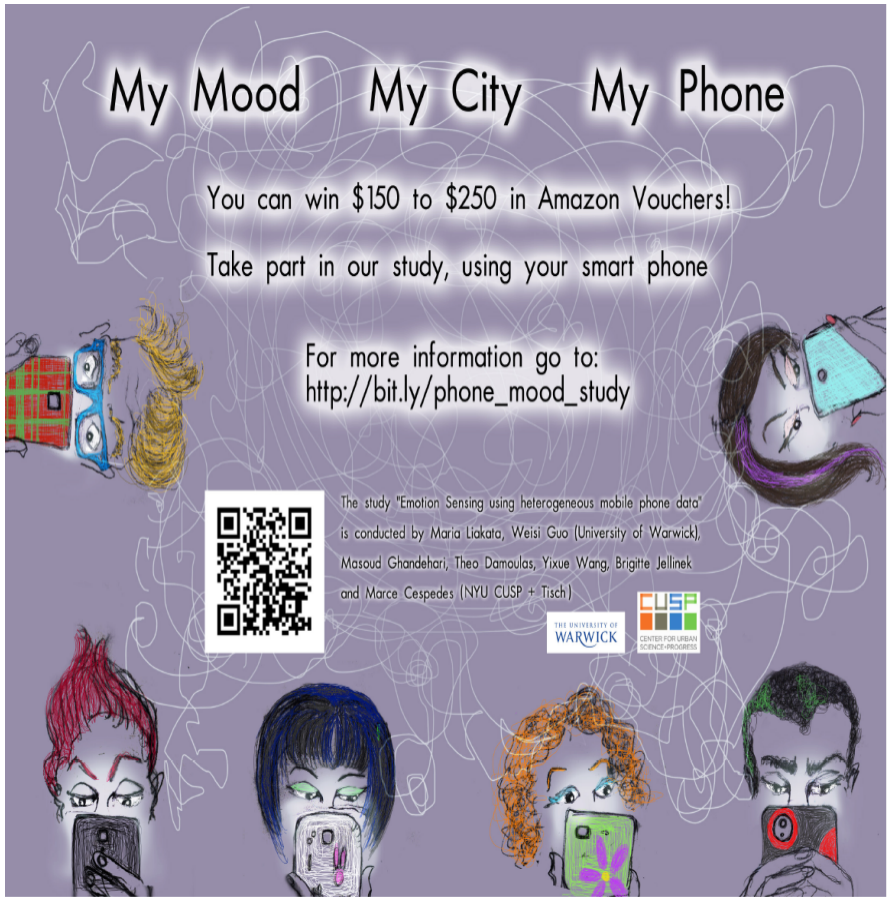
- TEXT: posts/messages (>110K)
  - Facebook, Twitter, Google+
  - SMS
- DA: mobile phone (~42GB)
  - Device Analyzer [Wagner et al., 2013]
  - location, wifi, calls…



Dataset

29 users, ~4 months

- TEXT: posts/messages (>110K)
  - Facebook, Twitter, Google+
  - SMS
- DA: mobile phone (~42GB)
  - Device Analyzer [Wagner et al., 2013]
  - location, wifi, calls…
- Target: {positive, negative, wellbeing} (2,436)
  - Daily self-assessments
  - PANAS [Watson et al., 1988]
  - WEMWBS [Tennant et al., 2007]



Task Formulation

Target: {positive, negative, wellbeing}

Features: TEXT/DA over the past 24h

Task Formulation

Target: {positive, negative, wellbeing}

Features: TEXT/DA over the past 24h

TEXT: merged (SMS, social media)
- ngrams
- lexicons
- word embeddings
- topics
- others (count-based)

DA: hour windows (1, 6, 12, 18, 24)
- locations
- wifi
- calls
- others (headphones, mode, charging, brightness, etc.)

- Missing DA data: past 6h mean
- 19 users / 1,438 instances

Experiments

Setup	user-agnostic 5-fold CV
Evaluation	RMSE, R^2
Transformation	feature & per-user normalisation
Baselines	LR, LASSO, RF, SVR

Experiments

Setup	user-agnostic 5-fold CV
Evaluation	RMSE, Rˆ2
Transformation	feature & per-user normalisation
Baselines	LR, LASSO, RF, SVR
MKL	one kernel per modality [Sonnenburg et al., 2006]

$$K(a,b) = \sum_{k=1}^{|K|} w_k K_k(a,b)$$

$$f(x) = \sum_{i=1}^{|N|} \alpha_i K(x, x_i) + b$$

Results

		Positive				Negative				Well-being			
		+User Norm		–User Norm		+User Norm		–User Norm		+User Norm		–User Norm	
		R^2	ϵ	R^2	ϵ	R^2	ϵ	R^2	ϵ	R^2	ϵ	R^2	ϵ
DA	LASSO	$_{n,n}$.31	8.24	$_{s}$.35	7.99	$_{n,n}$.11	6.71	$_{s}$.22	6.25	$_{n,n}$.30	10.51	$_{s}$.35	10.15
	RF	$_{s,s}$.69	5.55	$_{s}$.64	5.95	$_{s,s}$.43	5.38	_.40	5.49	$_{s,s}$.75	6.33	$_{s}$.67	7.18
	SVR	$_{n,n}$.58	6.38	$_{n}$.60	6.27	$_{n,n}$.35	5.74	$_{n}$.36	5.69	$_{n,n}$.62	7.80	$_{n}$.62	7.77
	MKL	$_{n,n}$.61	6.15	_.59	6.36	$_{n,n}$.38	5.60	$_{n}$.33	5.82	$_{n,s}$.65	7.43	$_{n}$.62	7.80
TEXT	LASSO	$_{n,n}$.53	6.80	$_{n}$.06	9.59	$_{n,n}$.23	6.23	$_{n}$.02	7.02	$_{n,n}$.55	8.46	$_{n}$.10	11.96
	RF	$_{-,s}$.70	5.42	$_{n}$.13	9.22	$_{-,s}$.45	5.26	$_{n}$.07	6.85	$_{s,s}$.74	6.36	$_{s}$.21	11.19
	SVR	$_{n,n}$.60	6.27	$_{n}$.11	9.31	$_{n,n}$.32	5.87	$_{n}$.06	6.88	$_{n,n}$.62	7.72	$_{n}$.19	11.30
	MKL	$_{n,n}$.62	6.08	$_{n}$.14	9.16	$_{n,n}$.36	5.69	_.06	6.89	$_{n,n}$.65	7.43	$_{n}$.22	11.12
ALL	LASSO	$_{n,n}$.49	7.07	$_{n}$.31	8.20	$_{n,n}$.18	6.41	$_{n}$.20	6.33	$_{n,n}$.54	8.52	$_{n}$.38	9.92
	RF	$_{-,s}$.71	5.31	$_{n}$.63	6.00	$_{-,s}$.46	5.20	$_{s}$.40	5.51	$_{n,s}$.76	6.23	_.68	7.12
	SVR	$_{n,n}$.60	6.27	$_{n}$.55	6.62	$_{n,n}$.34	5.76	$_{n}$.31	5.88	$_{n,n}$.62	7.75	$_{n}$.58	8.17
	MKL	$_{n,n}$.65	5.84	$_{n}$.61	6.14	$_{n,n}$.41	5.45	$_{n}$.36	5.67	$_{n,n}$.68	7.12	$_{n}$.64	7.58

Results

		Positive				Negative				Well-being			
		+User Norm		−User Norm		+User Norm		−User Norm		+User Norm		−User Norm	
		R^2	ϵ	R^2	ϵ	R^2	ϵ	R^2	ϵ	R^2	ϵ	R^2	ϵ
DA	LASSO	$_{n,n}$.31	8.24	$_{s}$.35	7.99	$_{n,n}$.11	6.71	$_{s}$.22	6.25	$_{n,n}$.30	10.51	$_{s}$.35	10.15
	RF	$_{s,s}$.69	5.55	$_{s}$.64	5.95	$_{s,s}$.43	5.38	$_{}$.40	5.49	$_{s,s}$.75	6.33	$_{s}$.67	7.18
	SVR	$_{n,n}$.58	6.38	$_{n}$.60	6.27	$_{n,n}$.35	5.74	$_{n}$.36	5.69	$_{n,n}$.62	7.80	$_{n}$.62	7.77
	MKL	$_{n,n}$.61	6.15	$_{}$.59	6.36	$_{n,n}$.38	5.60	$_{n}$.33	5.82	$_{n,s}$.65	7.43	$_{n}$.62	7.80
TEXT	LASSO	$_{n,n}$.53	6.80	$_{n}$.06	9.59	$_{n,n}$.23	6.23	$_{n}$.02	7.02	$_{n,n}$.55	8.46	$_{n}$.10	11.96
	RF	$_{,s}$.70	5.42	$_{n}$.13	9.22	$_{,s}$.45	5.26	$_{n}$.07	6.85	$_{s,s}$.74	6.36	$_{s}$.21	11.19
	SVR	$_{n,n}$.60	6.27	$_{n}$.11	9.31	$_{n,n}$.32	5.87	$_{n}$.06	6.88	$_{n,n}$.62	7.72	$_{n}$.19	11.30
	MKL	$_{n,n}$.62	6.08	$_{n}$.14	9.16	$_{n,n}$.36	5.69	$_{}$.06	6.89	$_{n,n}$.65	7.43	$_{n}$.22	11.12
ALL	LASSO	$_{n,n}$.49	7.07	$_{n}$.31	8.20	$_{n,n}$.18	6.41	$_{n}$.20	6.33	$_{n,n}$.54	8.52	$_{n}$.38	9.92
	RF	$_{,s}$.71	5.31	$_{n}$.63	6.00	$_{,s}$.46	5.20	$_{s}$.40	5.51	$_{n,s}$.76	6.23	$_{}$.68	7.12
	SVR	$_{n,n}$.60	6.27	$_{n}$.55	6.62	$_{n,n}$.34	5.76	$_{n}$.31	5.88	$_{n,n}$.62	7.75	$_{n}$.58	8.17
	MKL	$_{n,n}$.65	5.84	$_{n}$.61	6.14	$_{n,n}$.41	5.45	$_{n}$.36	5.67	$_{n,n}$.68	7.12	$_{n}$.64	7.58

Results

		Positive				Negative				Well-being			
		+User Norm		–User Norm		+User Norm		–User Norm		+User Norm		–User Norm	
		R^2	ϵ	R^2	ϵ	R^2	ϵ	R^2	ϵ	R^2	ϵ	R^2	ϵ
DA	LASSO	$_{n,n}$.31	8.24	$_{s}$.35	7.99	$_{n,n}$.11	6.71	$_{s}$.22	6.25	$_{n,n}$.30	10.51	$_{s}$.35	10.15
	RF	$_{s,s}$.69	5.55	$_{s}$.64	5.95	$_{s,s}$.43	5.38	$_{}$.40	5.49	$_{s,s}$.75	6.33	$_{s}$.67	7.18
	SVR	$_{n,n}$.58	6.38	$_{n}$.60	6.27	$_{n,n}$.35	5.74	$_{n}$.36	5.69	$_{n,n}$.62	7.80	$_{n}$.62	7.77
	MKL	$_{n,n}$.61	6.15	$_{}$.59	6.36	$_{n,n}$.38	5.60	$_{n}$.33	5.82	$_{n,s}$.65	7.43	$_{n}$.62	7.80
TEXT	LASSO	$_{n,n}$.53	6.80	$_{n}$.06	9.59	$_{n,n}$.23	6.23	$_{n}$.02	7.02	$_{n,n}$.55	8.46	$_{n}$.10	11.96
	RF	$_{,s}$.70	5.42	$_{n}$.13	9.22	$_{,s}$.45	5.26	$_{n}$.07	6.85	$_{s,s}$.74	6.36	$_{s}$.21	11.19
	SVR	$_{n,n}$.60	6.27	$_{n}$.11	9.31	$_{n,n}$.32	5.87	$_{n}$.06	6.88	$_{n,n}$.62	7.72	$_{n}$.19	11.30
	MKL	$_{n,n}$.62	6.08	$_{n}$.14	9.16	$_{n,n}$.36	5.69	$_{}$.06	6.89	$_{n,n}$.65	7.43	$_{n}$.22	11.12
ALL	LASSO	$_{n,n}$.49	7.07	$_{n}$.31	8.20	$_{n,n}$.18	6.41	$_{n}$.20	6.33	$_{n,n}$.54	8.52	$_{n}$.38	9.92
	RF	$_{,s}$.71	5.31	$_{n}$.63	6.00	$_{,s}$.46	5.20	$_{s}$.40	5.51	$_{n,s}$.76	6.23	$_{}$.68	7.12
	SVR	$_{n,n}$.60	6.27	$_{n}$.55	6.62	$_{n,n}$.34	5.76	$_{n}$.31	5.88	$_{n,n}$.62	7.75	$_{n}$.58	8.17
	MKL	$_{n,n}$.65	5.84	$_{n}$.61	6.14	$_{n,n}$.41	5.45	$_{n}$.36	5.67	$_{n,n}$.68	7.12	$_{n}$.64	7.58

Results

		Positive				Negative				Well-being			
		+User Norm		–User Norm		+User Norm		–User Norm		+User Norm		–User Norm	
		R^2	ϵ	R^2	ϵ	R^2	ϵ	R^2	ϵ	R^2	ϵ	R^2	ϵ
DA	LASSO	$_{n,n}$.31	8.24	$_{s}$.35	7.99	$_{n,n}$.11	6.71	$_{s}$.22	6.25	$_{n,n}$.30	10.51	$_{s}$.35	10.15
	RF	$_{s,s}$.69	5.55	$_{s}$.64	5.95	$_{s,s}$.43	5.38	$_{}$.40	5.49	$_{s,s}$.75	6.33	$_{s}$.67	7.18
	SVR	$_{n,n}$.58	6.38	$_{n}$.60	6.27	$_{n,n}$.35	5.74	$_{n}$.36	5.69	$_{n,n}$.62	7.80	$_{n}$.62	7.77
	MKL	$_{n,n}$.61	6.15	$_{}$.59	6.36	$_{n,n}$.38	5.60	$_{n}$.33	5.82	$_{n,s}$.65	7.43	$_{n}$.62	7.80
TEXT	LASSO	$_{n,n}$.53	6.80	$_{n}$.06	9.59	$_{n,n}$.23	6.23	$_{n}$.02	7.02	$_{n,n}$.55	8.46	$_{n}$.10	11.96
	RF	$_{,s}$.70	5.42	$_{n}$.13	9.22	$_{,s}$.45	5.26	$_{n}$.07	6.85	$_{s,s}$.74	6.36	$_{s}$.21	11.19
	SVR	$_{n,n}$.60	6.27	$_{n}$.11	9.31	$_{n,n}$.32	5.87	$_{n}$.06	6.88	$_{n,n}$.62	7.72	$_{n}$.19	11.30
	MKL	$_{n,n}$.62	6.08	$_{n}$.14	9.16	$_{n,n}$.36	5.69	$_{}$.06	6.89	$_{n,n}$.65	7.43	$_{n}$.22	11.12
ALL	LASSO	$_{n,n}$.49	7.07	$_{n}$.31	8.20	$_{n,n}$.18	6.41	$_{n}$.20	6.33	$_{n,n}$.54	8.52	$_{n}$.38	9.92
	RF	$_{,s}$.71	5.31	$_{n}$.63	6.00	$_{,s}$.46	5.20	$_{s}$.40	5.51	$_{n,s}$.76	6.23	$_{}$.68	7.12
	SVR	$_{n,n}$.60	6.27	$_{n}$.55	6.62	$_{n,n}$.34	5.76	$_{n}$.31	5.88	$_{n,n}$.62	7.75	$_{n}$.58	8.17
	MKL	$_{n,n}$.65	5.84	$_{n}$.61	6.14	$_{n,n}$.41	5.45	$_{n}$.36	5.67	$_{n,n}$.68	7.12	$_{n}$.64	7.58

Results

		Positive				Negative				Well-being			
		+User Norm		−User Norm		+User Norm		−User Norm		+User Norm		−User Norm	
		R^2	ϵ	R^2	ϵ	R^2	ϵ	R^2	ϵ	R^2	ϵ	R^2	ϵ
DA	LASSO	$_{n,n}$.31	8.24	$_{s}$.35	7.99	$_{n,n}$.11	6.71	$_{s}$.22	6.25	$_{n,n}$.30	10.51	$_{s}$.35	10.15
	RF	$_{s,s}$.69	5.55	$_{s}$.64	5.95	$_{s,s}$.43	5.38	$_{}$.40	5.49	$_{s,s}$.75	6.33	$_{s}$.67	7.18
	SVR	$_{n,n}$.58	6.38	$_{n}$.60	6.27	$_{n,n}$.35	5.74	$_{n}$.36	5.69	$_{n,n}$.62	7.80	$_{n}$.62	7.77
	MKL	$_{n,n}$.61	6.15	$_{}$.59	6.36	$_{n,n}$.38	5.60	$_{n}$.33	5.82	$_{n,s}$.65	7.43	$_{n}$.62	7.80
TEXT	LASSO	$_{n,n}$.53	6.80	$_{n}$.06	9.59	$_{n,n}$.23	6.23	$_{n}$.02	7.02	$_{n,n}$.55	8.46	$_{n}$.10	11.96
	RF	$_{,s}$.70	5.42	$_{n}$.13	9.22	$_{,s}$.45	5.26	$_{n}$.07	6.85	$_{s,s}$.74	6.36	$_{s}$.21	11.19
	SVR	$_{n,n}$.60	6.27	$_{n}$.11	9.31	$_{n,n}$.32	5.87	$_{n}$.06	6.88	$_{n,n}$.62	7.72	$_{n}$.19	11.30
	MKL	$_{n,n}$.62	6.08	$_{n}$.14	9.16	$_{n,n}$.36	5.69	$_{}$.06	6.89	$_{n,n}$.65	7.43	$_{n}$.22	11.12
ALL	LASSO	$_{n,n}$.49	7.07	$_{n}$.31	8.20	$_{n,n}$.18	6.41	$_{n}$.20	6.33	$_{n,n}$.54	8.52	$_{n}$.38	9.92
	RF	$_{,s}$.71	5.31	$_{n}$.63	6.00	$_{,s}$.46	5.20	$_{s}$.40	5.51	$_{n,s}$.76	6.23	$_{}$.68	7.12
	SVR	$_{n,n}$.60	6.27	$_{n}$.55	6.62	$_{n,n}$.34	5.76	$_{n}$.31	5.88	$_{n,n}$.62	7.75	$_{n}$.58	8.17
	MKL	$_{n,n}$.65	5.84	$_{n}$.61	6.14	$_{n,n}$.41	5.45	$_{n}$.36	5.67	$_{n,n}$.68	7.12	$_{n}$.64	7.58

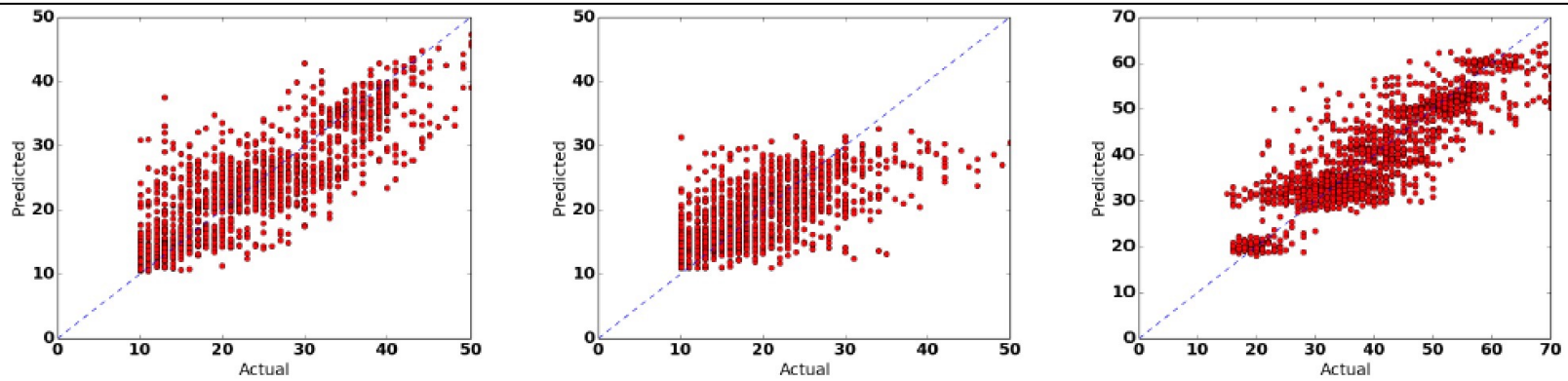
Results

		Positive				Negative				Well-being			
		+User Norm		−User Norm		+User Norm		−User Norm		+User Norm		−User Norm	
		R^2	ϵ	R^2	ϵ	R^2	ϵ	R^2	ϵ	R^2	ϵ	R^2	ϵ
DA	LASSO	$_{n,n}$.31	8.24	$_{s}$.35	7.99	$_{n,n}$.11	6.71	$_{s}$.22	6.25	$_{n,n}$.30	10.51	$_{s}$.35	10.15
	RF	$_{s,s}$.69	5.55	$_{s}$.64	5.95	$_{s,s}$.43	5.38	$_{}$.40	5.49	$_{s,s}$.75	6.33	$_{s}$.67	7.18
	SVR	$_{n,n}$.58	6.38	$_{n}$.60	6.27	$_{n,n}$.35	5.74	$_{n}$.36	5.69	$_{n,n}$.62	7.80	$_{n}$.62	7.77
	MKL	$_{n,n}$.61	6.15	$_{}$.59	6.36	$_{n,n}$.38	5.60	$_{n}$.33	5.82	$_{n,s}$.65	7.43	$_{n}$.62	7.80
TEXT	LASSO	$_{n,n}$.53	6.80	$_{n}$.06	9.59	$_{n,n}$.23	6.23	$_{n}$.02	7.02	$_{n,n}$.55	8.46	$_{n}$.10	11.96
	RF	$_{,s}$.70	5.42	$_{n}$.13	9.22	$_{,s}$.45	5.26	$_{n}$.07	6.85	$_{s,s}$.74	6.36	$_{s}$.21	11.19
	SVR	$_{n,n}$.60	6.27	$_{n}$.11	9.31	$_{n,n}$.32	5.87	$_{n}$.06	6.88	$_{n,n}$.62	7.72	$_{n}$.19	11.30
	MKL	$_{n,n}$.62	6.08	$_{n}$.14	9.16	$_{n,n}$.36	5.69	$_{}$.06	6.89	$_{n,n}$.65	7.43	$_{n}$.22	11.12
ALL	LASSO	$_{n,n}$.49	7.07	$_{n}$.31	8.20	$_{n,n}$.18	6.41	$_{n}$.20	6.33	$_{n,n}$.54	8.52	$_{n}$.38	9.92
	RF	$_{,s}$.71	5.31	$_{n}$.63	6.00	$_{,s}$.46	5.20	$_{s}$.40	5.51	$_{n,s}$.76	6.23	$_{}$.68	7.12
	SVR	$_{n,n}$.60	6.27	$_{n}$.55	6.62	$_{n,n}$.34	5.76	$_{n}$.31	5.88	$_{n,n}$.62	7.75	$_{n}$.58	8.17
	MKL	$_{n,n}$.65	5.84	$_{n}$.61	6.14	$_{n,n}$.41	5.45	$_{n}$.36	5.67	$_{n,n}$.68	7.12	$_{n}$.64	7.58

Results

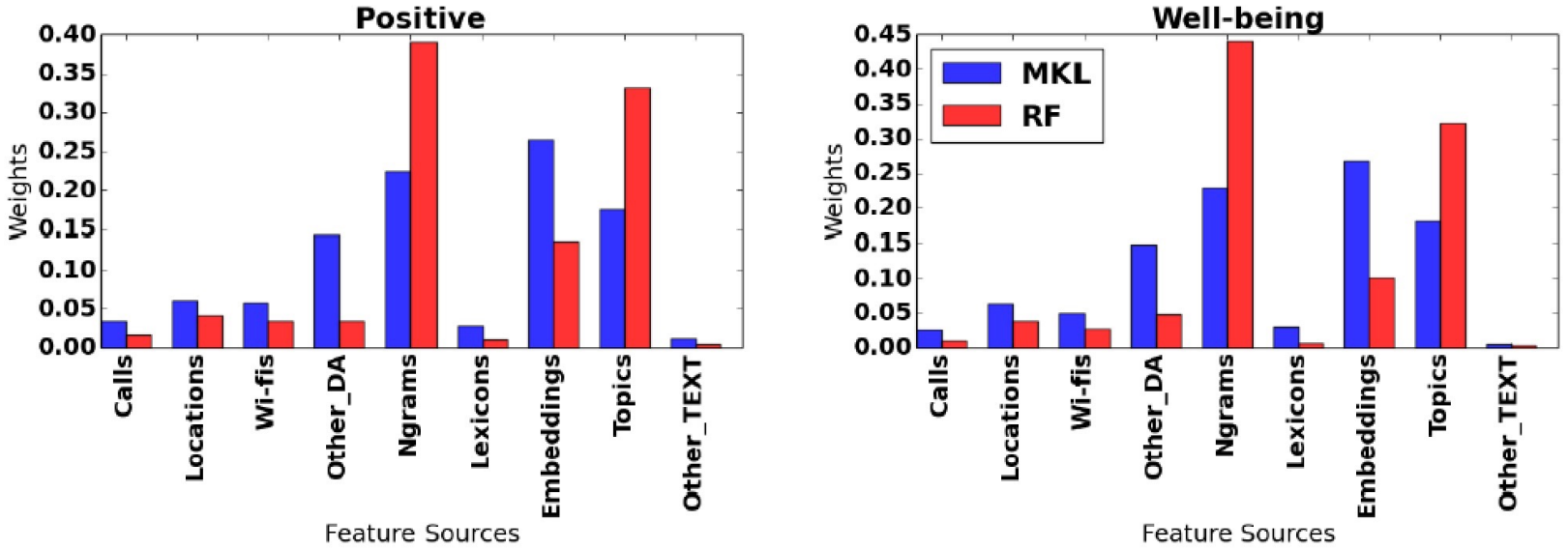
		Positive				Negative				Well-being			
		+User Norm		–User Norm		+User Norm		–User Norm		+User Norm		–User Norm	
		R^2	ϵ	R^2	ϵ	R^2	ϵ	R^2	ϵ	R^2	ϵ	R^2	ϵ
DA	LASSO	$_{n,n}$.31	8.24	$_{s}$.35	7.99	$_{n,n}$.11	6.71	$_{s}$.22	6.25	$_{n,n}$.30	10.51	$_{s}$.35	10.15
	RF	$_{s,s}$.69	5.55	$_{s}$.64	5.95	$_{s,s}$.43	5.38	_.40	5.49	$_{s,s}$.75	6.33	$_{s}$.67	7.18
	SVR	$_{n,n}$.58	6.38	$_{n}$.60	6.27	$_{n,n}$.35	5.74	$_{n}$.36	5.69	$_{n,n}$.62	7.80	$_{n}$.62	7.77
	MKL	$_{n,n}$.61	6.15	_.59	6.36	$_{n,n}$.38	5.60	$_{n}$.33	5.82	$_{n,s}$.65	7.43	$_{n}$.62	7.80
TEXT	LASSO	$_{n,n}$.53	6.80	$_{n}$.06	9.59	$_{n,n}$.23	6.23	$_{n}$.02	7.02	$_{n,n}$.55	8.46	$_{n}$.10	11.96
	RF	$_{-,s}$.70	5.42	$_{n}$.13	9.22	$_{-,s}$.45	5.26	$_{n}$.07	6.85	$_{s,s}$.74	6.36	$_{s}$.21	11.19
	SVR	$_{n,n}$.60	6.27	$_{n}$.11	9.31	$_{n,n}$.32	5.87	$_{n}$.06	6.88	$_{n,n}$.62	7.72	$_{n}$.19	11.30
	MKL	$_{n,n}$.62	6.08	$_{n}$.14	9.16	$_{n,n}$.36	5.69	_.06	6.89	$_{n,n}$.65	7.43	$_{n}$.22	11.12
ALL	LASSO	$_{n,n}$.49	7.07	$_{n}$.31	8.20	$_{n,n}$.18	6.41	$_{n}$.20	6.33	$_{n,n}$.54	8.52	$_{n}$.38	9.92
	RF	$_{-,s}$.71	5.31	$_{n}$.63	6.00	$_{-,s}$.46	5.20	$_{s}$.40	5.51	$_{n,s}$.76	6.23	_.68	7.12
	SVR	$_{n,n}$.60	6.27	$_{n}$.55	6.62	$_{n,n}$.34	5.76	$_{n}$.31	5.88	$_{n,n}$.62	7.75	$_{n}$.58	8.17
	MKL	$_{n,n}$.65	5.84	$_{n}$.61	6.14	$_{n,n}$.41	5.45	$_{n}$.36	5.67	$_{n,n}$.68	7.12	$_{n}$.64	7.58

Results



(a) Positive (b) Negative (c) Well-being

Figure 1: Actual VS Predicted charts for the best performing algorithm (RF) on the three targets.



Discussion

Summary

- Task: continuous sensing for mental health assessment
- Novel heterogeneous dataset
- MKL: heterogeneity
- Accuracy? VERY high!

Discussion

Summary

- Task: continuous sensing for mental health assessment
- Novel heterogeneous dataset
- MKL: heterogeneity
- Accuracy? VERY high!

Next steps

- Longitudinal modelling
- Handling of missing data
- Different types of validation

Reference: 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).

Discussion

Summary

- Task: continuous sensing for mental health assessment
- Novel heterogeneous dataset
- MKL: heterogeneity
- **Accuracy? VERY high!**

Next steps

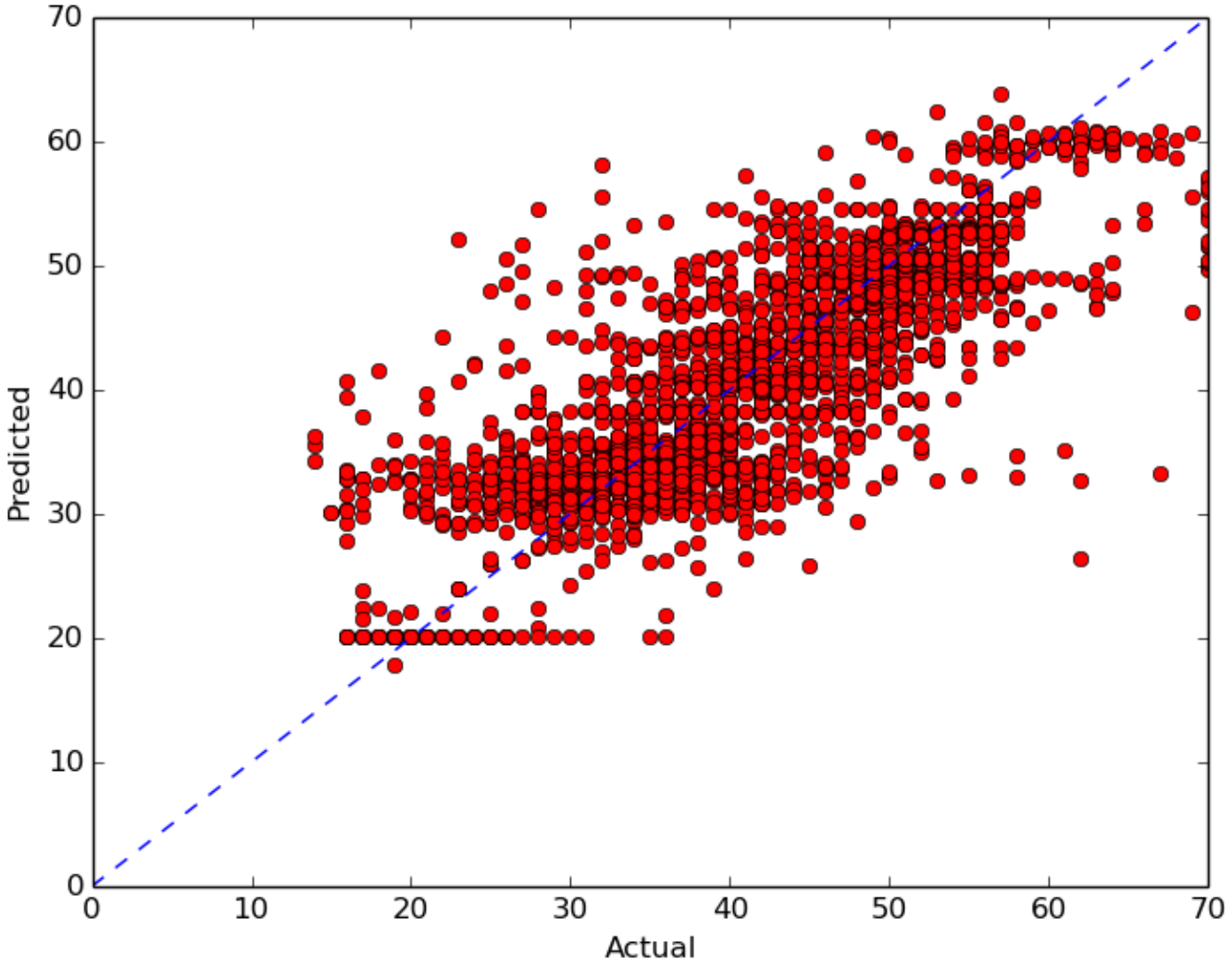
- Longitudinal modelling
- Handling of missing data
- **Different types of validation**

Reference: 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).

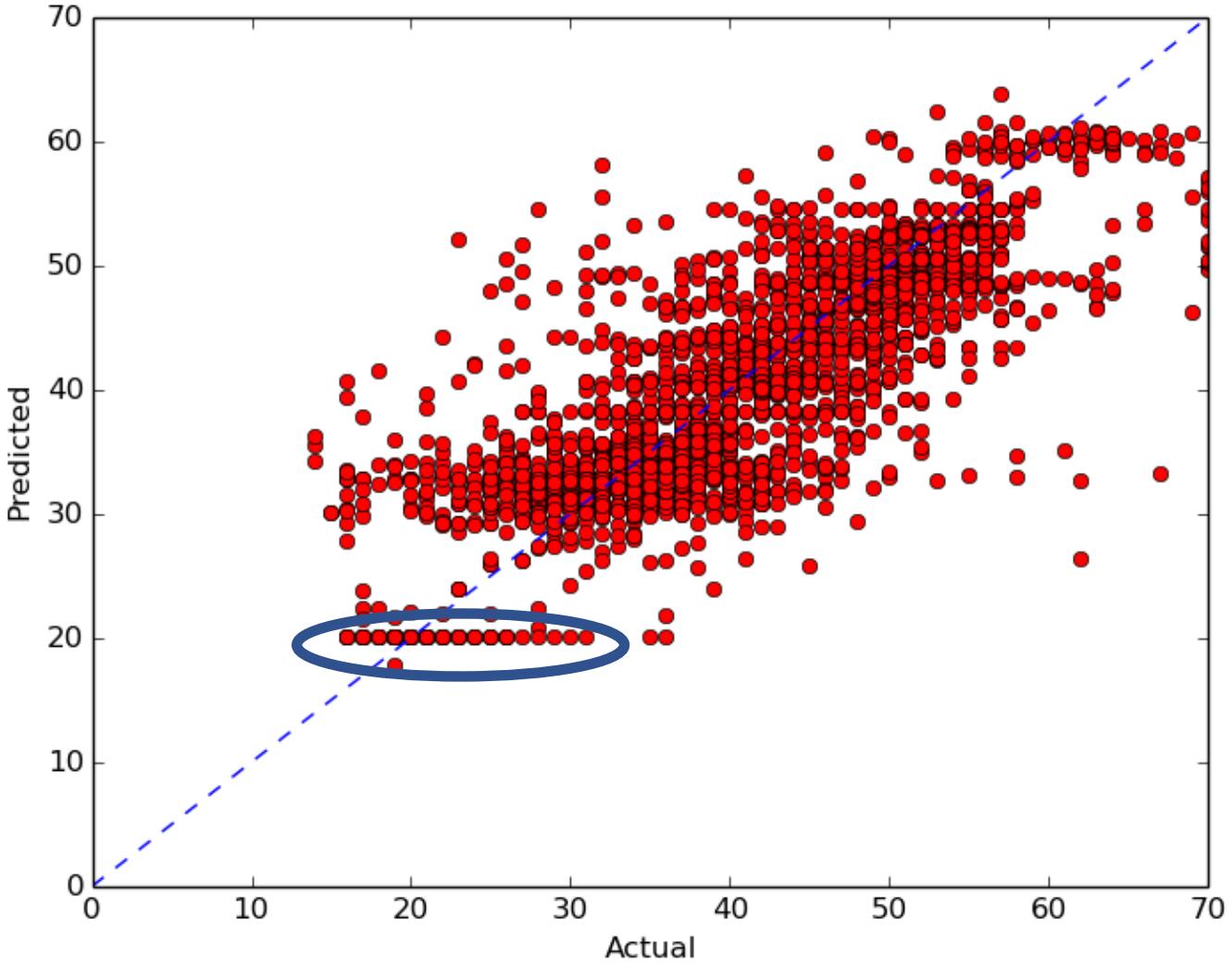
Overview

- **Problem Definition**
- **Part 1:** Assessing Mental Health with Smart Devices and Social Media
  - Dataset Collection
  - Experiments and Results
  - Discussion
- **Part 2:** Addressing Bias in Methodology and Evaluation
  - Real-world Challenges
  - Experiments and Results
- **Conclusion & Future Directions**

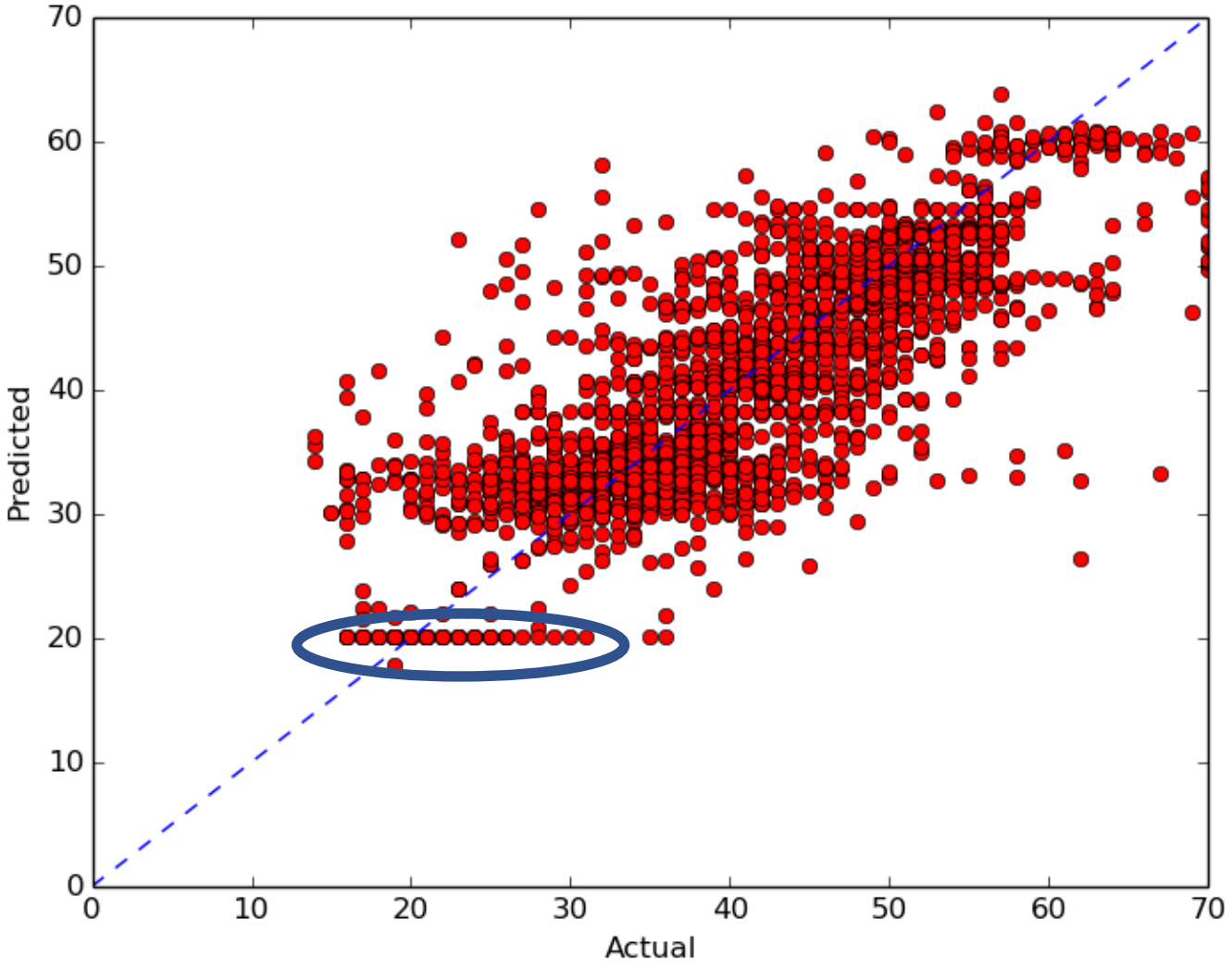
Revising Our Results



Revising Our Results

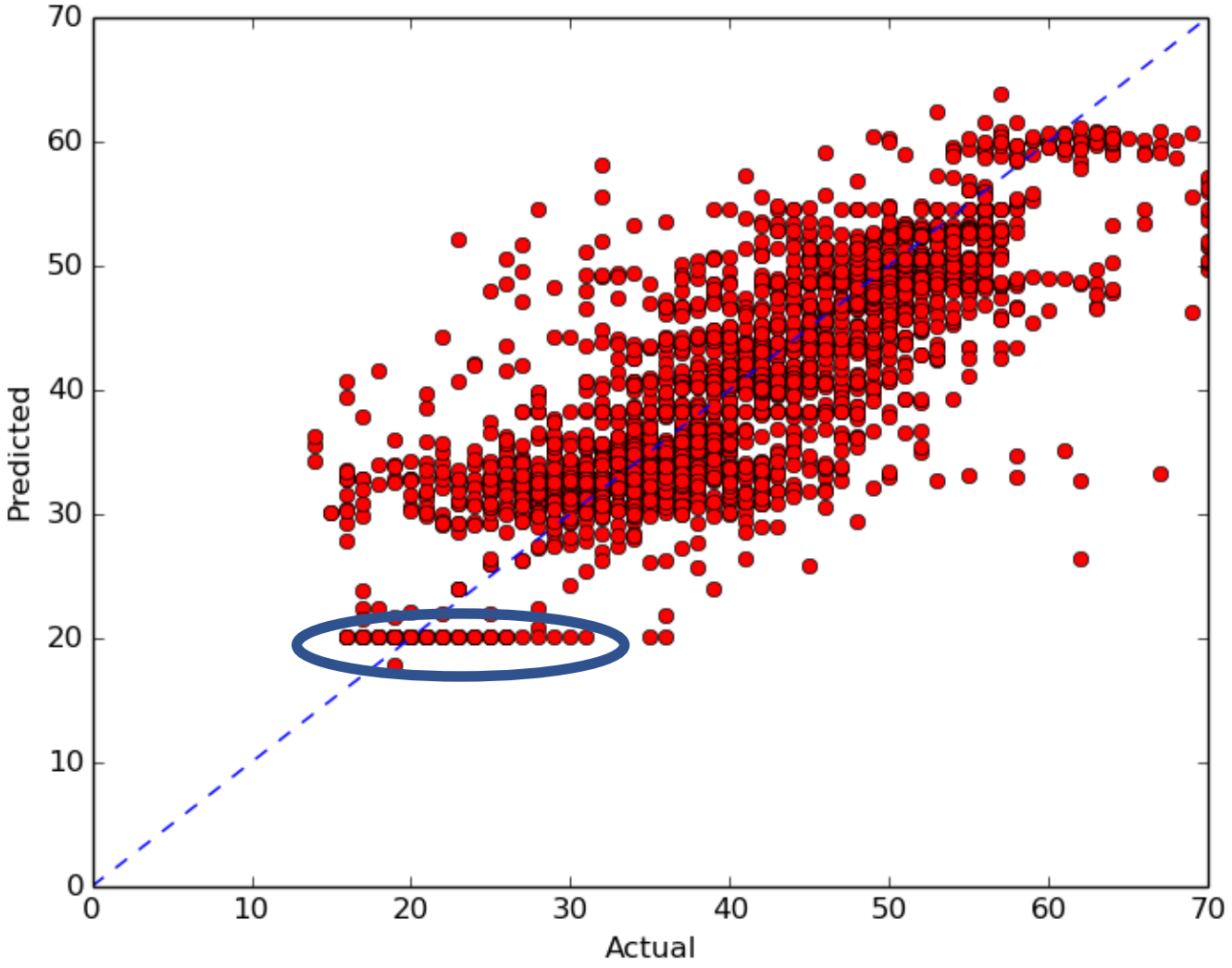


Revising Our Results

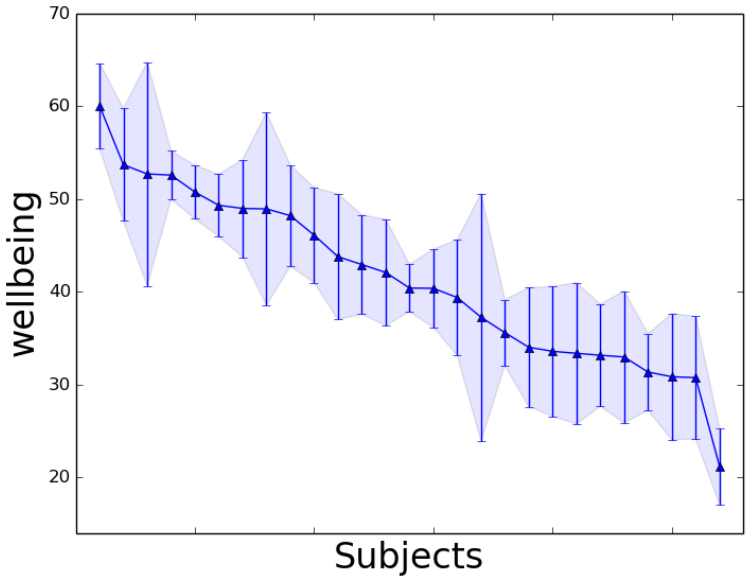


$$R^2 = 1 - \frac{\sum_i (y_i - y_{pred})^2}{\sum_i (y_i - y_{avg})^2}$$

Revising Our Results

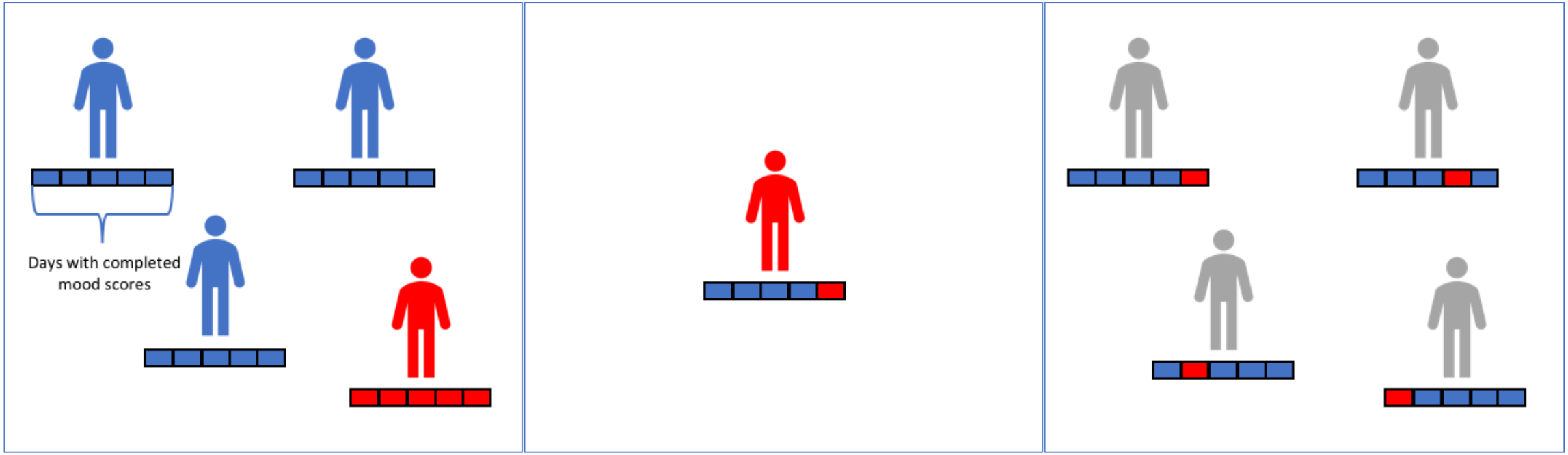


$$R^2 = 1 - \frac{\sum_i (y_i - y_{pred})^2}{\sum_i (y_i - y_{avg})^2}$$

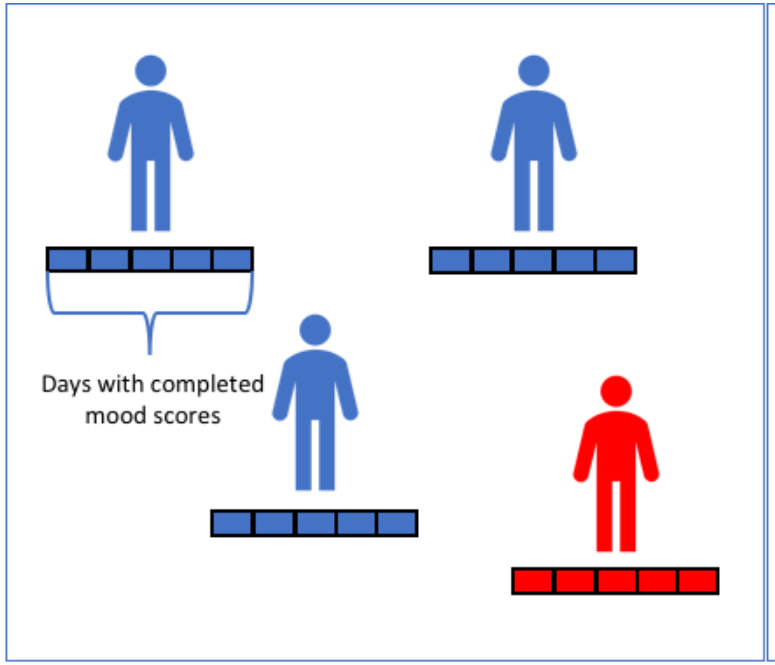


[DeMasi et al., 2017; Tsakalidis et al., 2018]

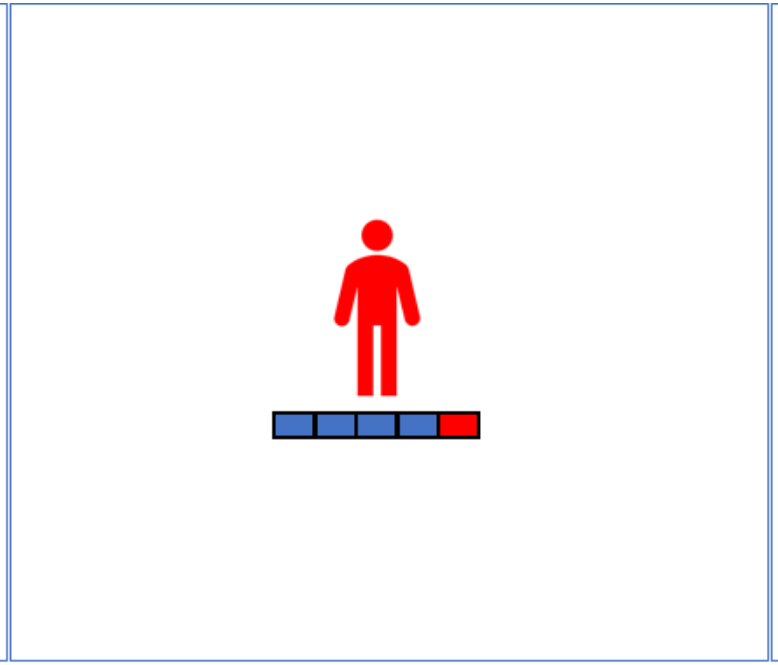
Types of Evaluation



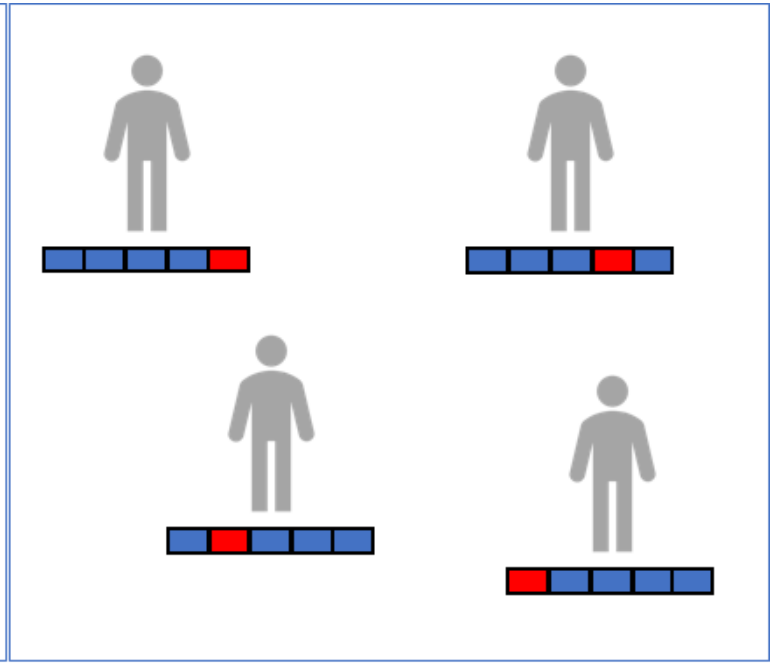
Types of Evaluation



LOUOCV



LOIOCV

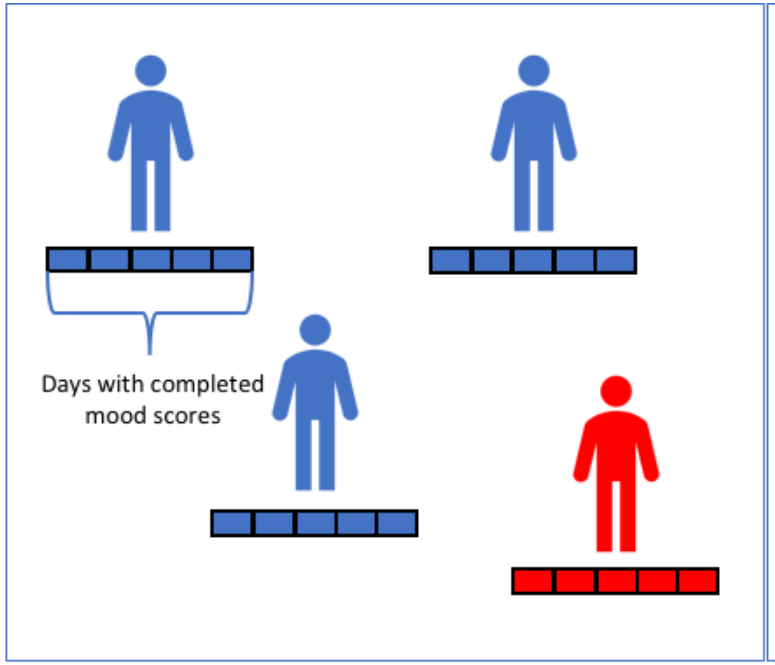


MIXED

Goal: Generalise to new users

Problem: few users…

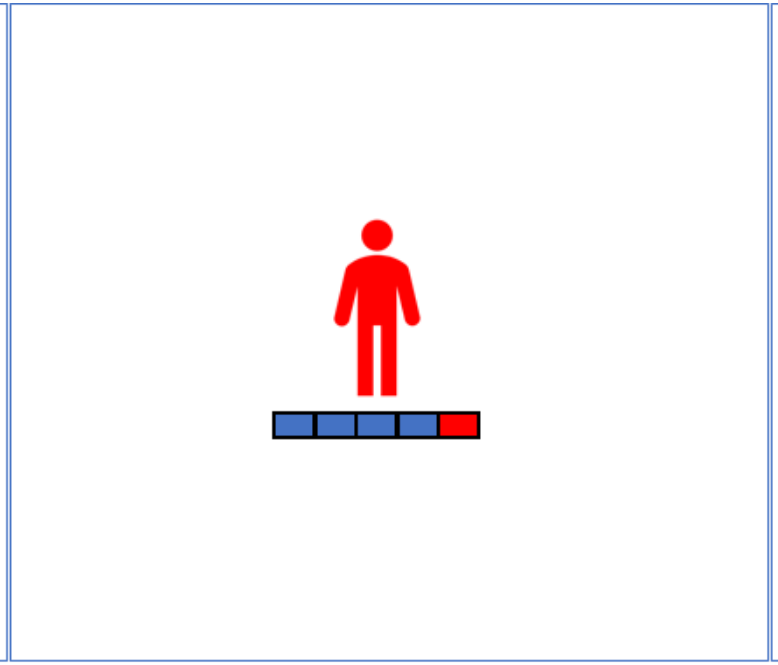
Types of Evaluation



LOUOCV

Goal: Generalise to new users

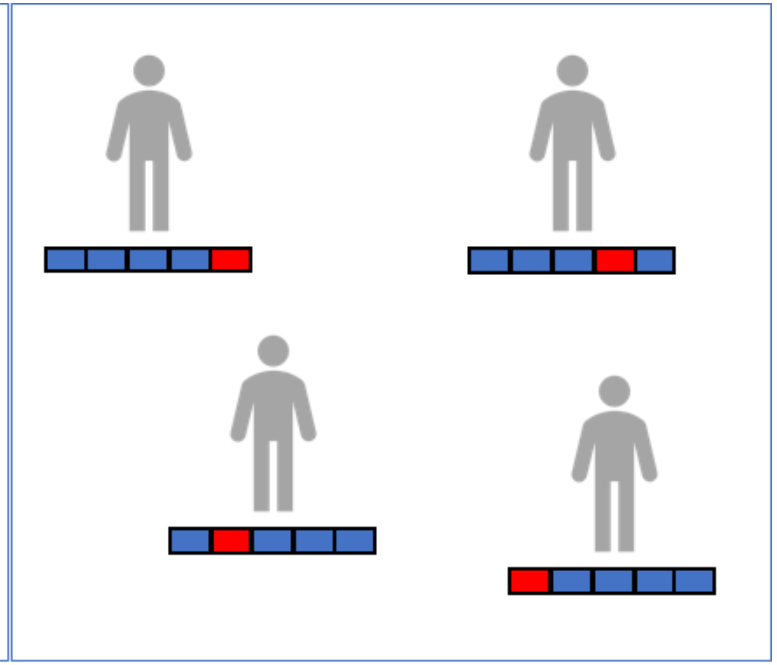
Problem: few users…



LOIOCV

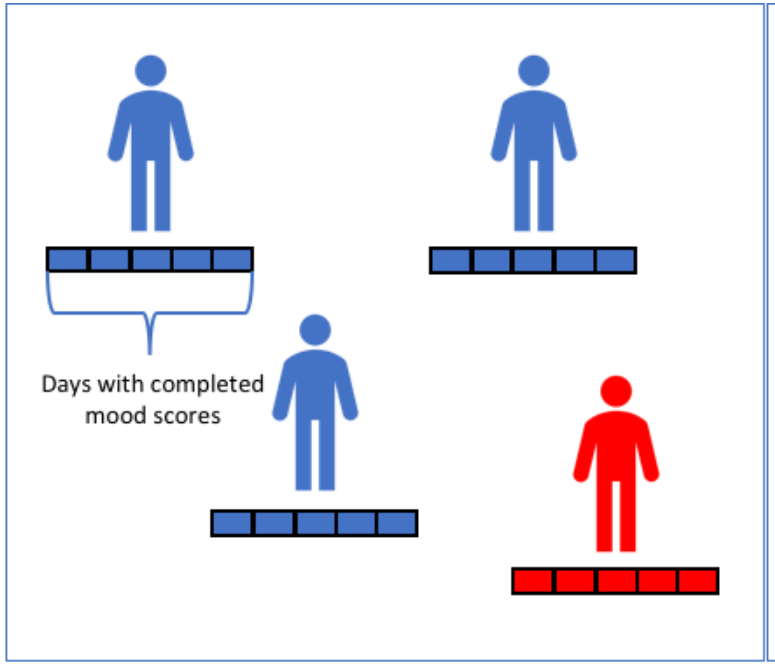
Goal: Personalised models

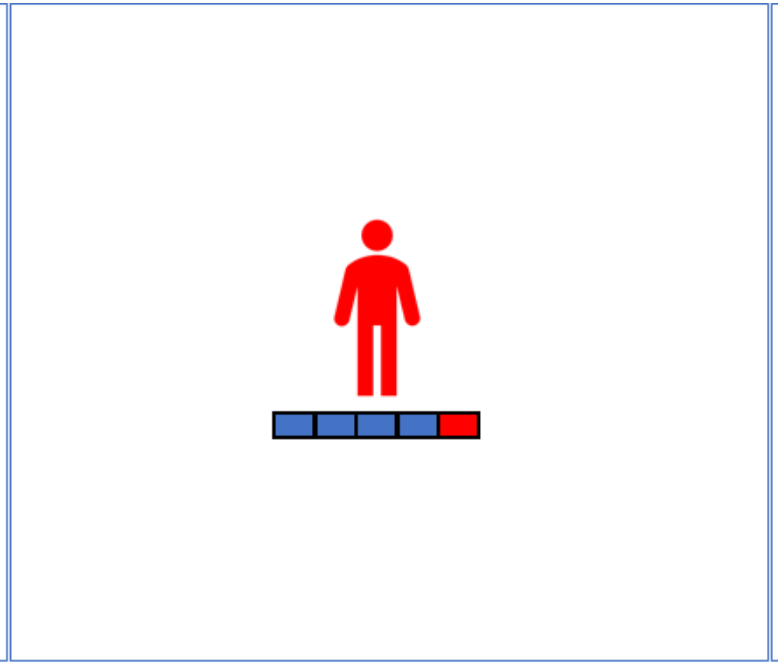
Problem: few instances…

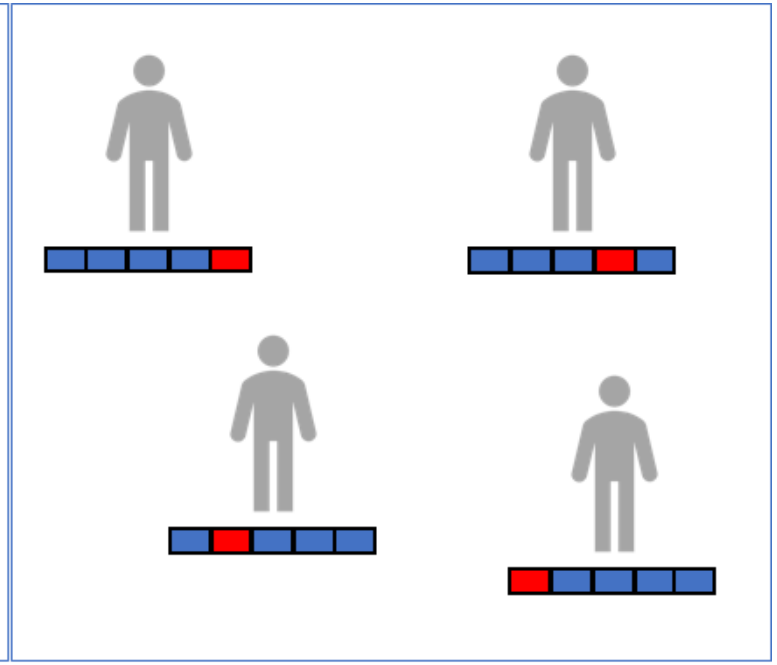


MIXED

Types of Evaluation







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

- Using past days' target scores as features
- Problems:
  - LOUOCV: 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

- Using past days' target scores as features
- Problems:
  - LOUOCV: new user?
  - LOIOCV: target score in test instance is used as a feature in another training example!

P2: Inferring test set labels

- Problem in LOIOCV:
  - Overlapping instances (input)
  - What if the target is also (temporally) correlated?

Problem Statement

P1: Training on past values of the target

- Using past days' target scores as features
- Problems:
  - LOUOCV: new user?
  - LOIOCV: target score in test instance is used as a feature in another training example!

P2: Inferring test set labels

- Problem in LOIOCV:
  - Overlapping instances (input)
  - What if the target is also (temporally) correlated?

P3: Predicting users instead of mood scores

- Problem in MIXED:
  - identify the user?

Problem Statement

Our goals:

- follow experimental/evaluation settings from leading work in the field
- alter to mimic a real-world setting
- compare model performance against naïve baselines
- propose directions for future research

Datasets and Features

- **Dataset 1** [Tsakalidis et al., 2016]
  - 27 subjects, ~4 months
  - **3 targets** (positive, negative, wellbeing)
  - **textual features**:

    ngrams, lexicons, word clusters, word embeddings, count-based

- **Dataset 2** [Wang et al., 2014]
  - 44 subjects, 10 weeks
  - **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., 2013] 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., 2013] 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., '13	
	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., '13	
	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

- **AVG**: best results; **-mood**: much worse!

Results: P1 (train on target values)

	positive		negative		wellbeing		stress		LiKamWa et al., '13	
	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

- **AVG**: best results; **-mood**: much worse!

LOIOCV

- Instances: overlapping time windows (3-days).
  What if our target is also correlated with respect to time?

P2: Inferring test set labels

[Canzian & Musolesi, 2015] extracted features from overlapping T_{HIST}

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

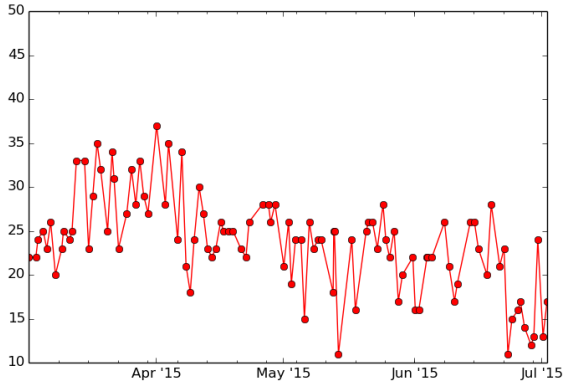
P2: Inferring test set labels

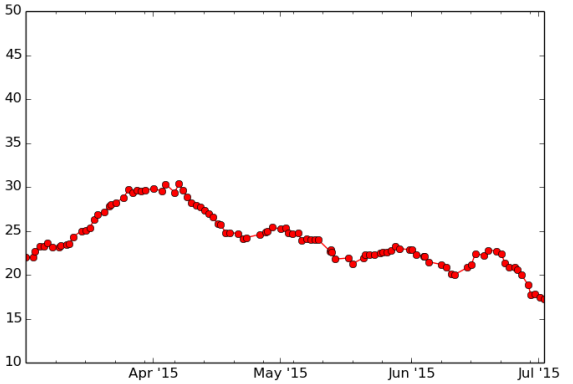
[Canzian & Musolesi, 2015] extracted features from overlapping T_{HIST}

- 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





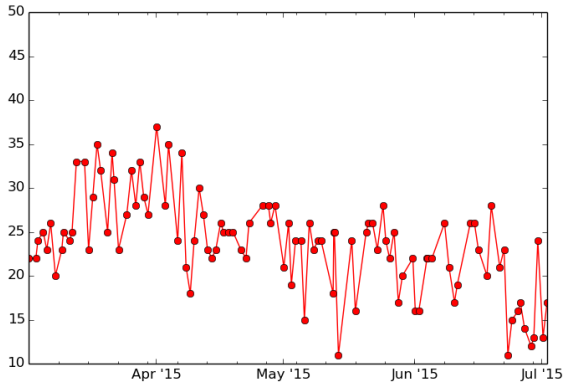
P2: Inferring test set labels

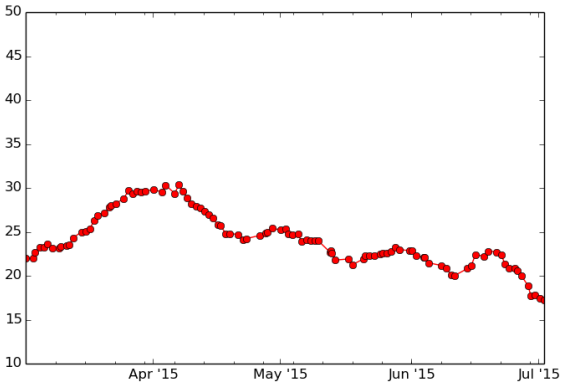
[Canzian & Musolesi, 2015] extracted features from overlapping T_{HIST}

- 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

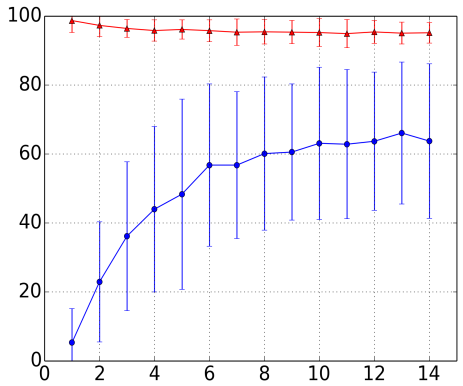
- LOIOCV
- binary (high/low) classification
- wider T_{HIST} => better results

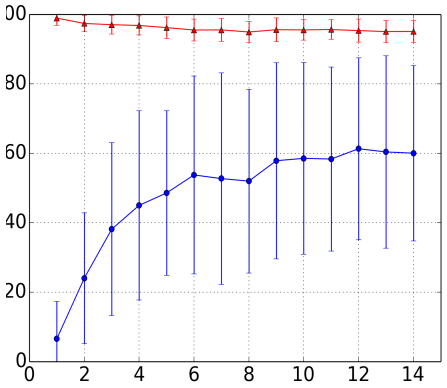
Results: P2 (inferring test set labels)

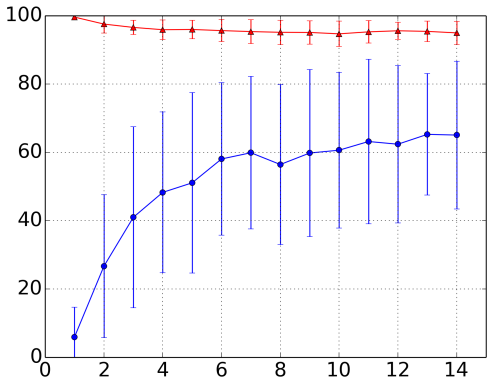
Results: P2 (inferring test set labels)

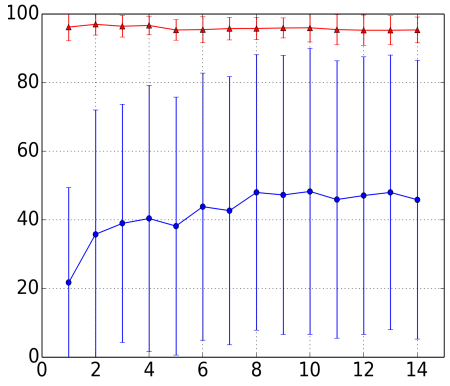
LOIOCV

- *Sensitivity:* increase with larger window size
- *Specificity:* stable, high





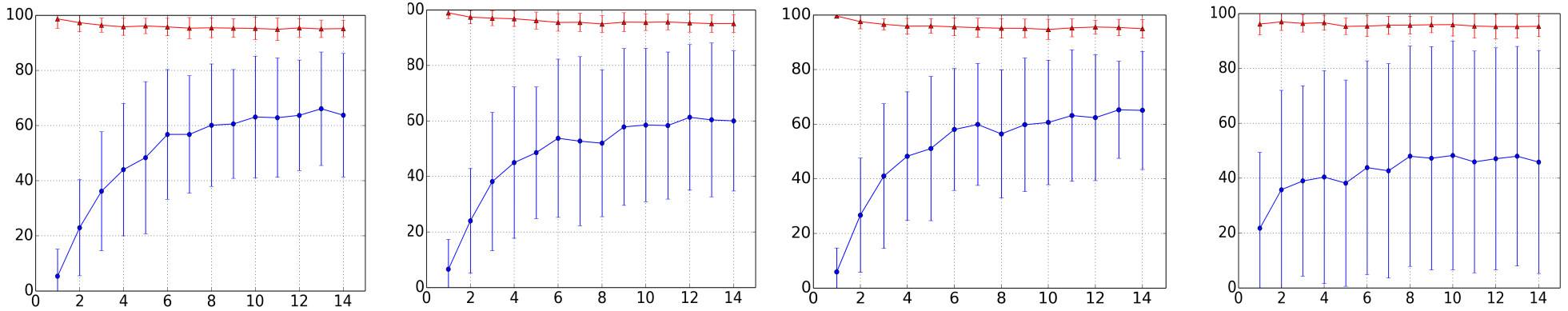




Results: P2 (inferring test set labels)

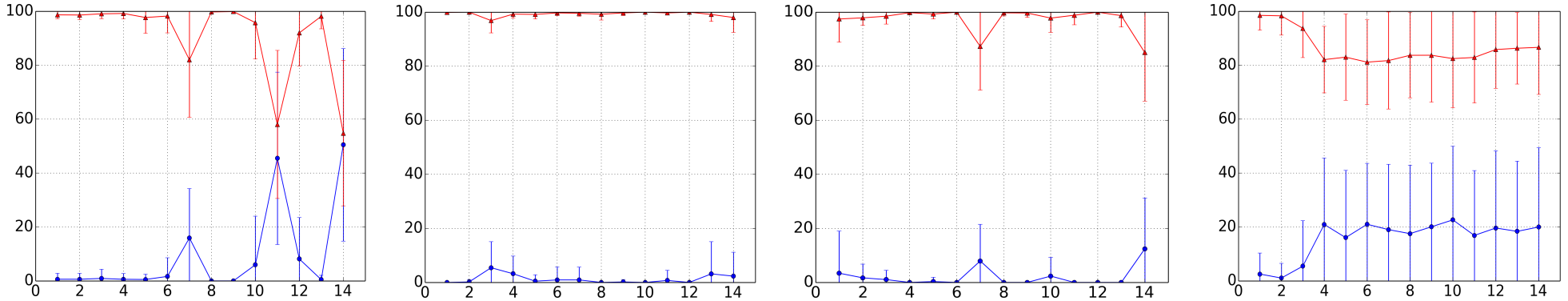
LOIOCV

- *Sensitivity:* increase with larger window size
- *Specificity:* stable, high



LOUOCV

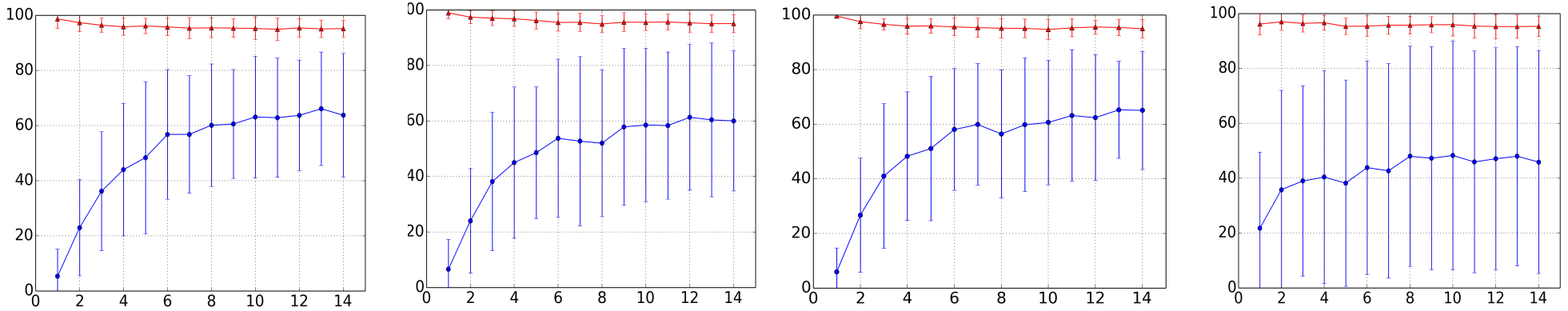
- Window size does not affect *sensitivity*
- Increase in sensitivity: sharp drops in *specificity*



Results: P2 (inferring test set labels)

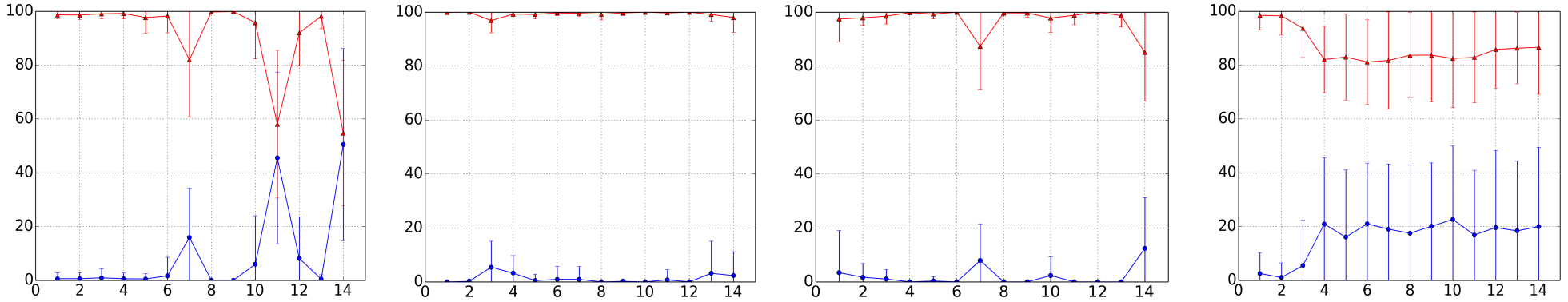
LOIOCV

- *Sensitivity:* increase with larger window size
- *Specificity:* stable, high



LOUOCV

- Window size does not affect *sensitivity*
- Increase in sensitivity: sharp drops in *specificity*



LOIOCV

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

FEAT:	64.02
DATE:	59.68
LAST:	67.37
RAND:	64.22

FEAT:	60.03
DATE:	62.75
LAST:	69.08
RAND:	60.88

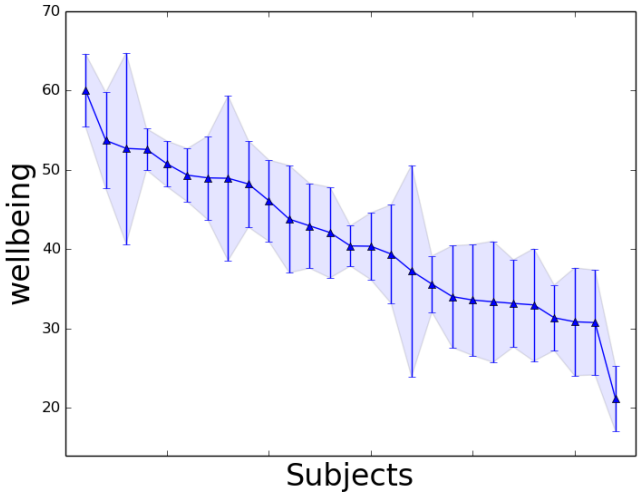
FEAT:	65.06
DATE:	63.29
LAST:	66.05
RAND:	64.87

FEAT:	45.86
DATE:	46.99
LAST:	58.20
RAND:	45.79

P3: Predicting users instead of mood scores

[Tsakalidis et al., 2016] evaluated regression models under MIXED

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



[Jaques et al., 2015] 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)

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

Experiment 1 (regression)

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)

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)

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)

MIXED: Effect of (per-user) feature normalisation

LOIOCV: Worse than the average mood predictor

LOUOCV: Results rather poor

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)

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)

UNIQ: Labelling instances across all users

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)

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)

UNIQ: Labelling instances across all users

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)

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)

UNIQ: Labelling instances across all users

PERS: Labelling instances on a per-user basis

Conclusions similar to E1

Overview

- **Problem Definition**
- **Part 1:** Assessing Mental Health with Smart Devices and Social Media
  - Dataset Collection
  - Experiments and Results
  - Discussion
- **Part 2:** Addressing Bias in Methodology and Evaluation
  - Real-world Challenges
  - Experiments and Results
- **Conclusion & Future Directions**

Future Directions

Issue: different users, different behaviour

- location1 for user1 != location1 for user2

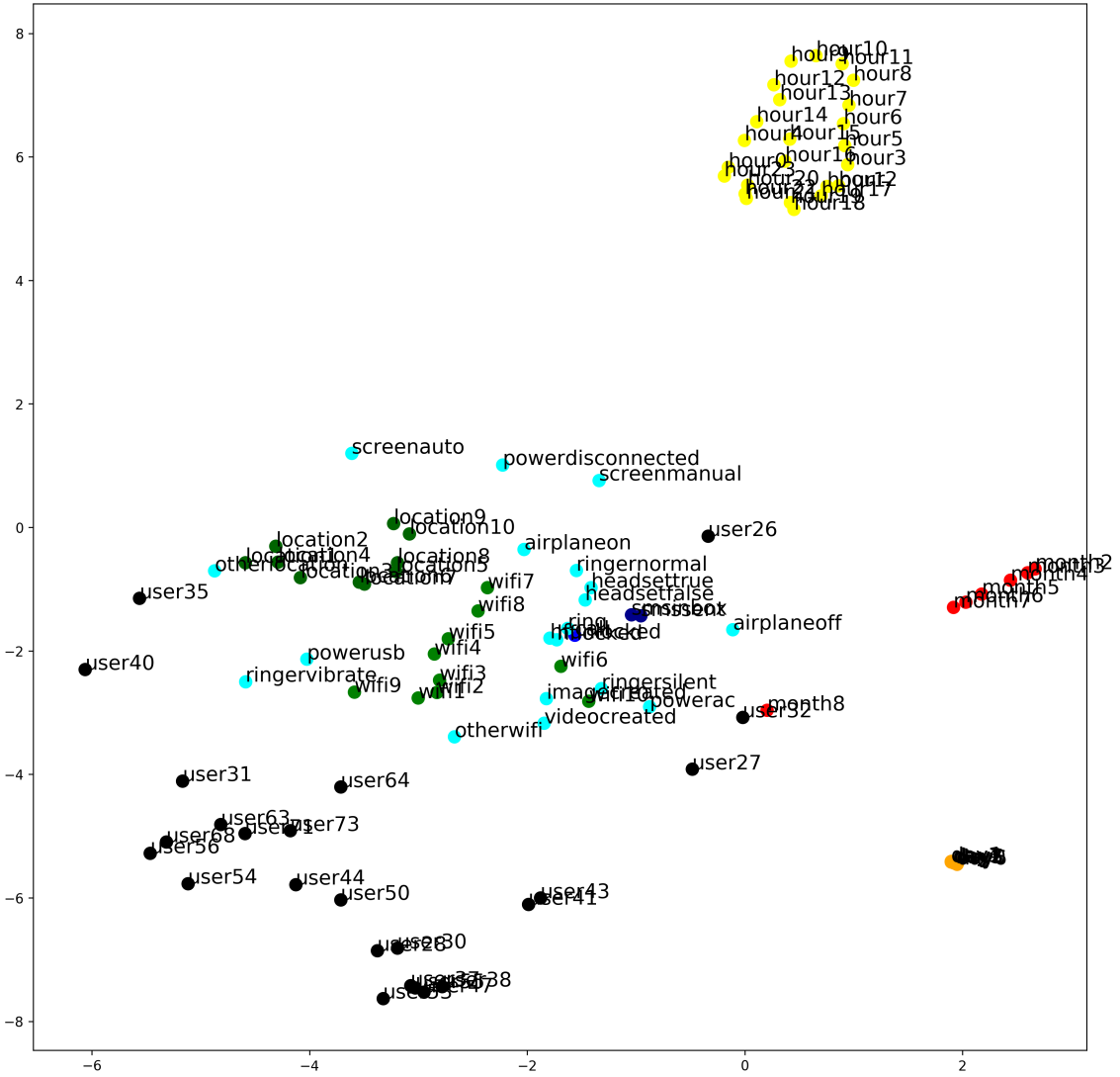
Proposal:

- latent feature representations [Mikolov et al., 2013]
- transfer learning
- demographics?
- larger datasets

Apply on strictly unbiased setting!

- LOUOCV
- LOIOCV

Future Directions (ongoing)



$$J = \log p_w(D = 1|f, h) + \sum_{\tilde{f} \sim Q_{\text{noise}}} \left[\log p_w(D = 0|\tilde{f}, h) \right]$$

Latent feature representations: incorporate the semantics of the features

Incorporation in ML models?

Conclusion

- Assessing mental health through smart devices & social media
  - Overview
  - Combining asynchronous and heterogeneous sources
  - Promising results (under certain evaluation setup…)

- Problem: SOTA does not follow a real-world setting!
  - Important for practitioners!
  - Conclusions reached might be wrong!

- Future Work
  - Highly important: evaluation setting
  - Larger datasets, latent features, transfer learning…

Special Thanks

Maria Liakata

Alexandra I. Cristea

Theo Damoulas

Weisi Guo

Brigitte Jellinek

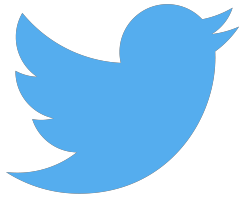
[1] Tsakalidis, A., Liakata, M., Damoulas, T., Jellinek, B., Guo, W. and Cristea, A., 2016. Combining Heterogeneous User Generated Data to Sense Well-being. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers* (pp. 3007-3018).

[2] Tsakalidis, A., Liakata, M., Damoulas, T. and Cristea, A.I., 2018. Can We Assess Mental Health through Social Media and Smart Devices? Addressing Bias in Methodology and Evaluation. *arXiv preprint arXiv:1807.07351*. (ECML PKDD 18/ADS, to appear)

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).

Thank you!

Any questions?



@adtsakal

atsakalidis@turing.ac.uk

References

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.

DeMasi, O., Kording, K. and Recht, B., 2017. Meaningless comparisons lead to false optimism in medical machine learning. *PloS one, 12(9)*, p.e0184604

Herrman, H., Saxena, S., Moodie, R. and World Health Organization, 2005. Promoting mental health: Concepts, emerging evidence, practice: A report of the World Health Organization, Department of Mental Health and Substance Abuse in collaboration with the Victorian Health Promotion Foundation and the University of Melbourne.

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.

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.

Mikolov, T., Sutskever, I., Chen, K., Corrado, G.S. and Dean, J., 2013. Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems* (pp. 3111-3119).

Sonnenburg, S., Rätsch, G., Schäfer, C. and Schölkopf, B., 2006. Large scale multiple kernel learning. *Journal of Machine Learning Research*, *7*(Jul), pp.1531-1565.

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.

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).

Tsakalidis, A., Liakata, M., Damoulas, T. and Cristea, A.I., 2018. Can We Assess Mental Health through Social Media and Smart Devices? Addressing Bias in Methodology and Evaluation. *arXiv preprint arXiv:1807.07351*.

Wagner, D.T., Rice, A. and Beresford, A.R., 2013, December. Device analyzer: Understanding smartphone usage. In *International Conference on Mobile and Ubiquitous Systems: Computing, Networking, and Services* (pp. 195-208). Springer, Cham.

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.

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.