Data Science for ESM data

What it is, Why it’s relevant, and How to do it

Drew Hendrickson

Cognitive Science & Artificial Intelligence

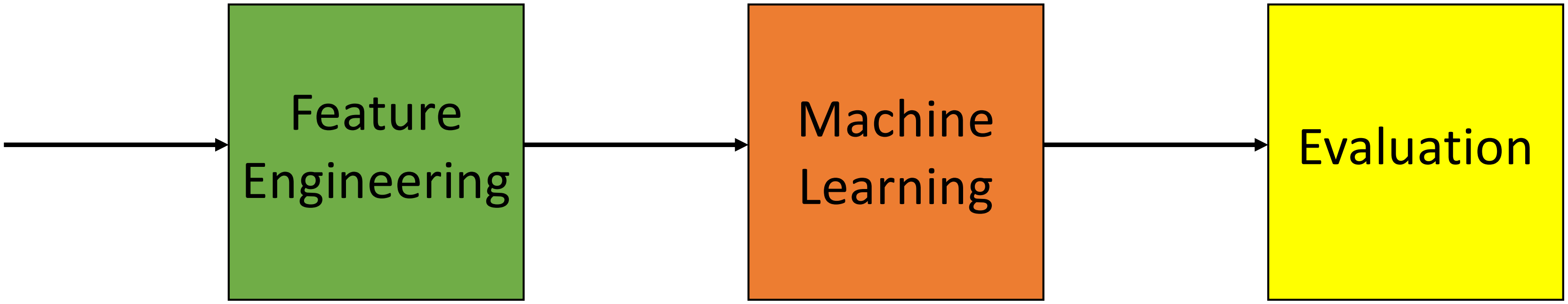
Tilburg University

TESC colloquium, October 4, 2022

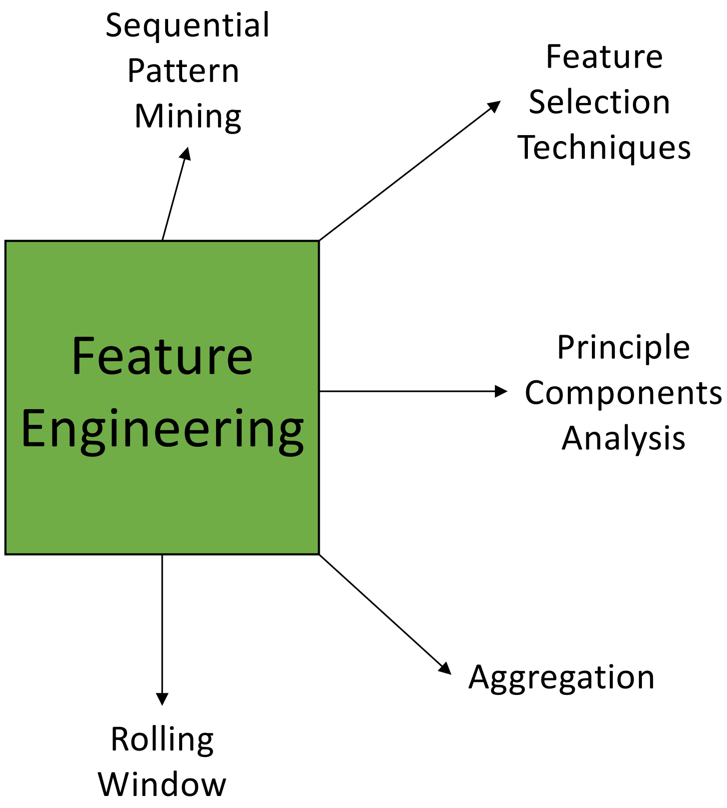
Outline

1. What are the components of the Data Science analysis pipeline?
2. What questions do Data Science analyses try to answer?
   - How are they different than inferential statistics or network analyses?
   - ASIDE: The uniqueness of ESM data
3. An example DS analysis: predicting stress in adolescents
   - Aalbers, et al., (in press, JMIR)

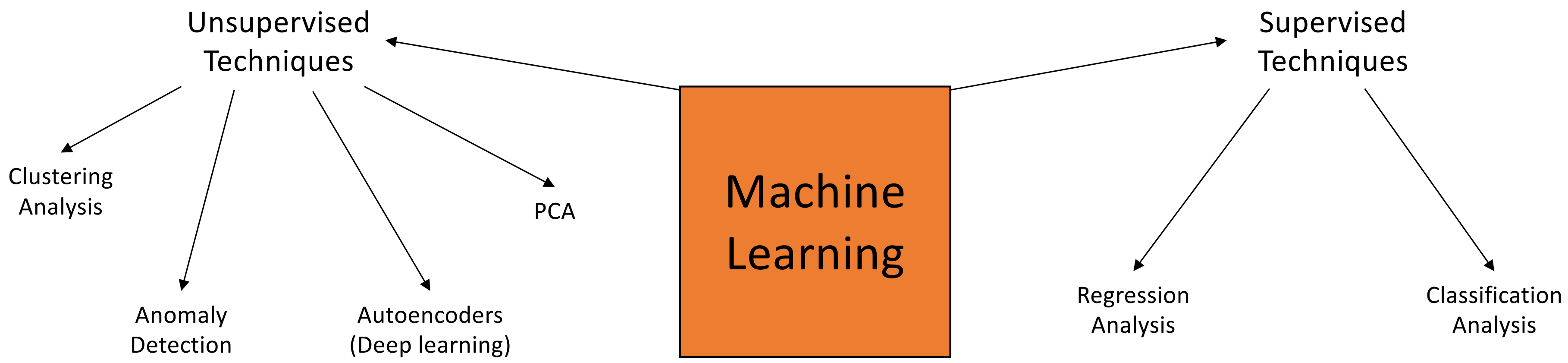
What is the Data Science pipeline?



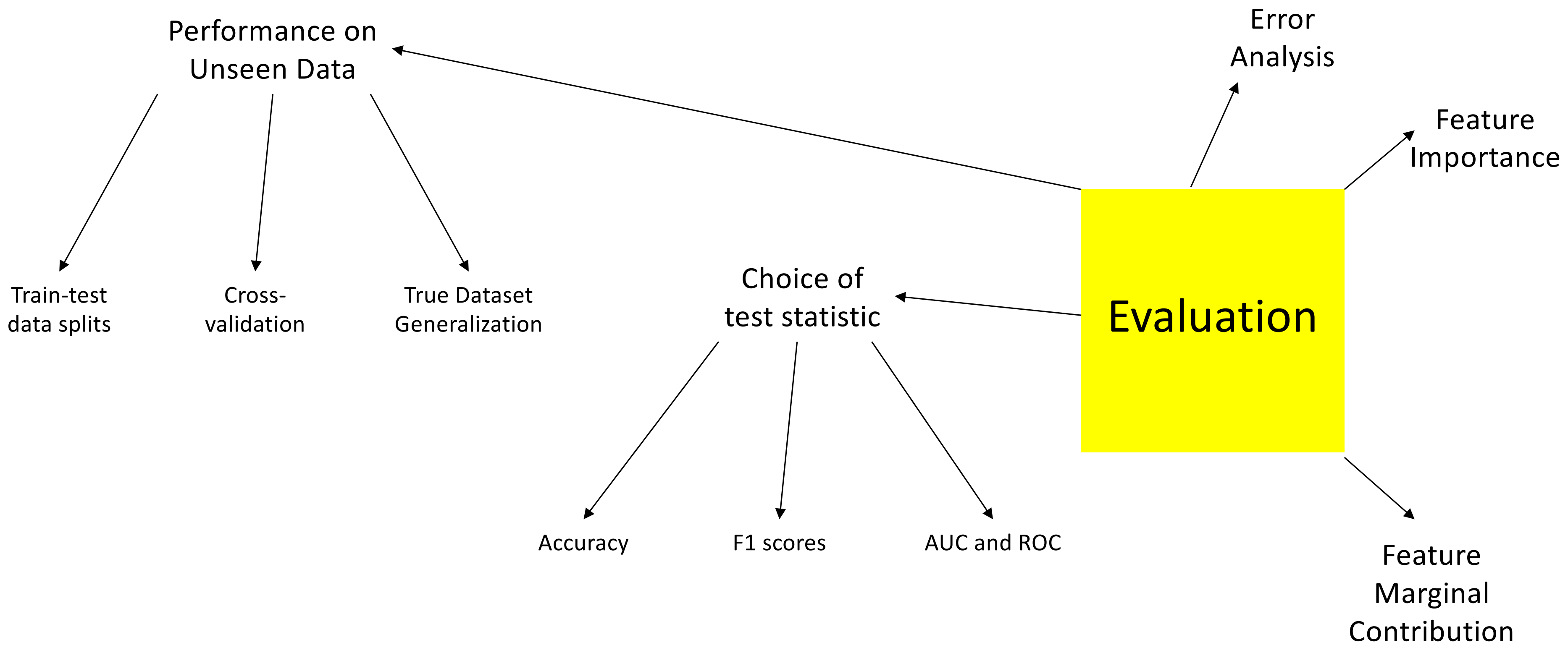
What is the Data Science pipeline?


Sequential Pattern Mining
Feature Selection Techniques
Feature Engineering
Principle Components Analysis
Aggregation
Rolling Window

What is the Data Science pipeline?



What is the Data Science pipeline?



What Q’s do Data Science analyses answer?

Consider an example: Linear Regression

What would a standard inferential statistical analysis tell us?

Model 1: `y ~ X`

Model 2: `y ~ X + Z`

```
anova(model_1, model_2)
```

Q: Is `Z` a significant predictor of `y`?

Q: Is there a sig. difference in `y` due to `Z`?

Follow-ups:

- Evaluate R^2 values
- Interpret beta weights

Consider an example: Linear Regression

What would a standard network analysis tell us?

For each `y` in `X`:

Model: $\mathtt{y}_{\mathtt{t}} \sim \mathtt{X}_{\mathtt{t-1}}$

Form matrix of beta weights as connections of measures from `t-1` to `t`

Q: What is the relationship between `X` values over time?

Follow-ups:

- Interpret beta weights (as partial correlations)
- Build nice networks that differentiate between sources of variance

Consider an example: Linear Regression

What would a standard data science analysis tell us?

Randomly split the data into a `training` set (70%) and `test` set (30%)

Model: $\mathtt{y}_{\mathtt{train}} \sim \mathtt{X}_{\mathtt{train}}$

$\mathtt{Accuracy = SSE(y_{test},\ Model(X_{test}))}$

Q: How well can we predict `y` using `X`?

Follow-ups:

- Evaluate R^2 values
- Feature importance
- Error analysis

What Q’s do Data Science analyses answer?

“How well can we predict y using X?”

IS: “Is Z a significant predictor of y?”

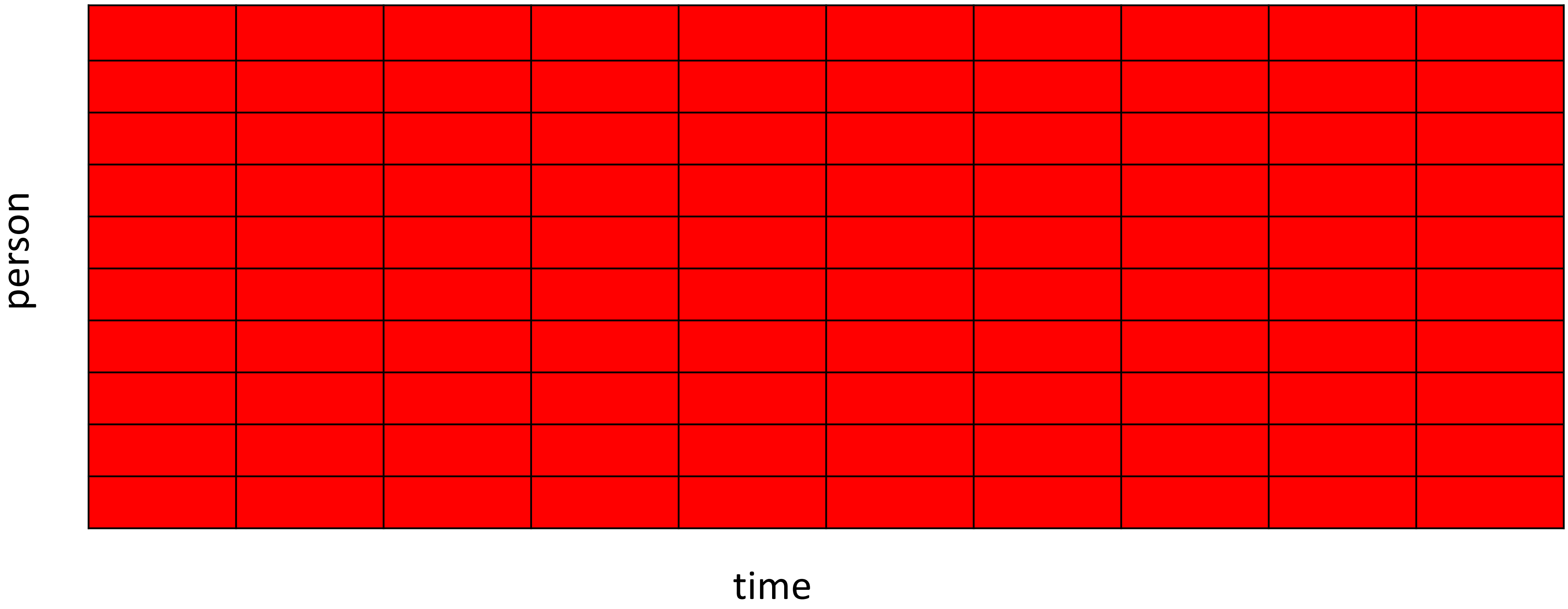
NA: “What is the relationship between X values over time? “

ASIDE: what is the ’correct’ test set for ESM?

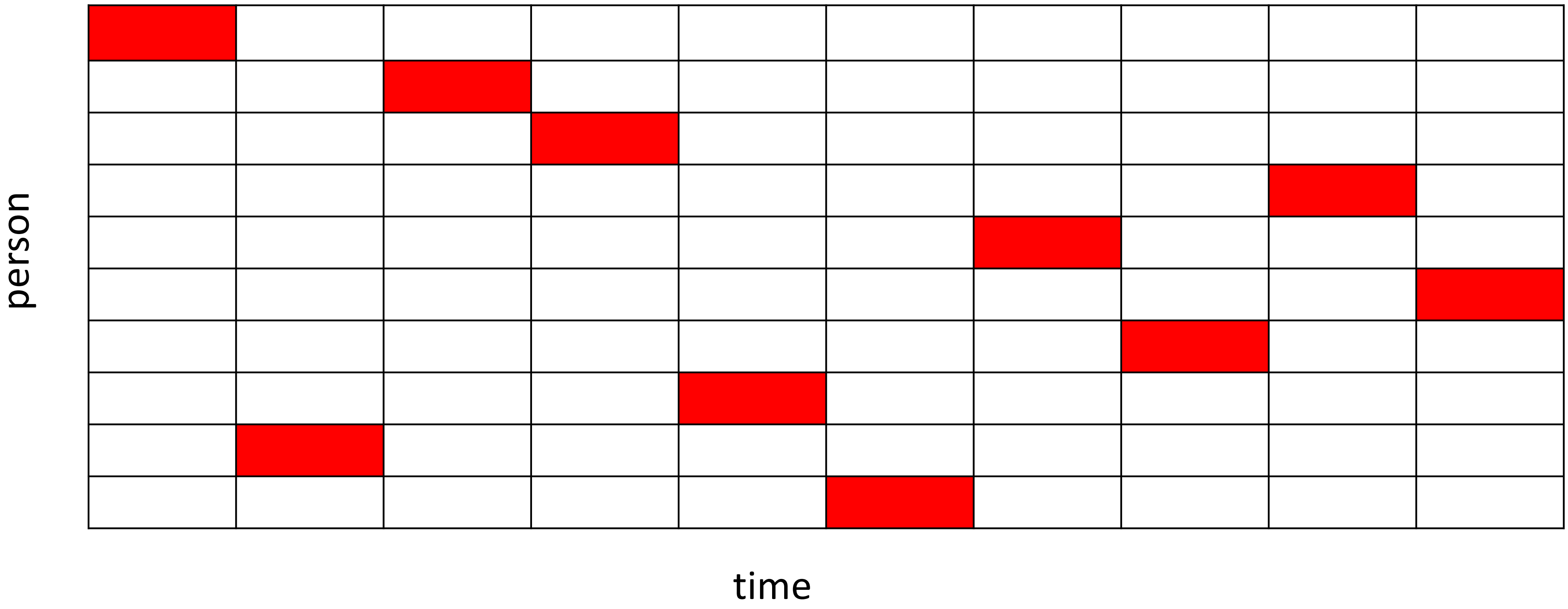
ASIDE: what is the ’correct’ test set for ESM?

person
time
"

ASIDE: what is the ’correct’ test set for ESM?

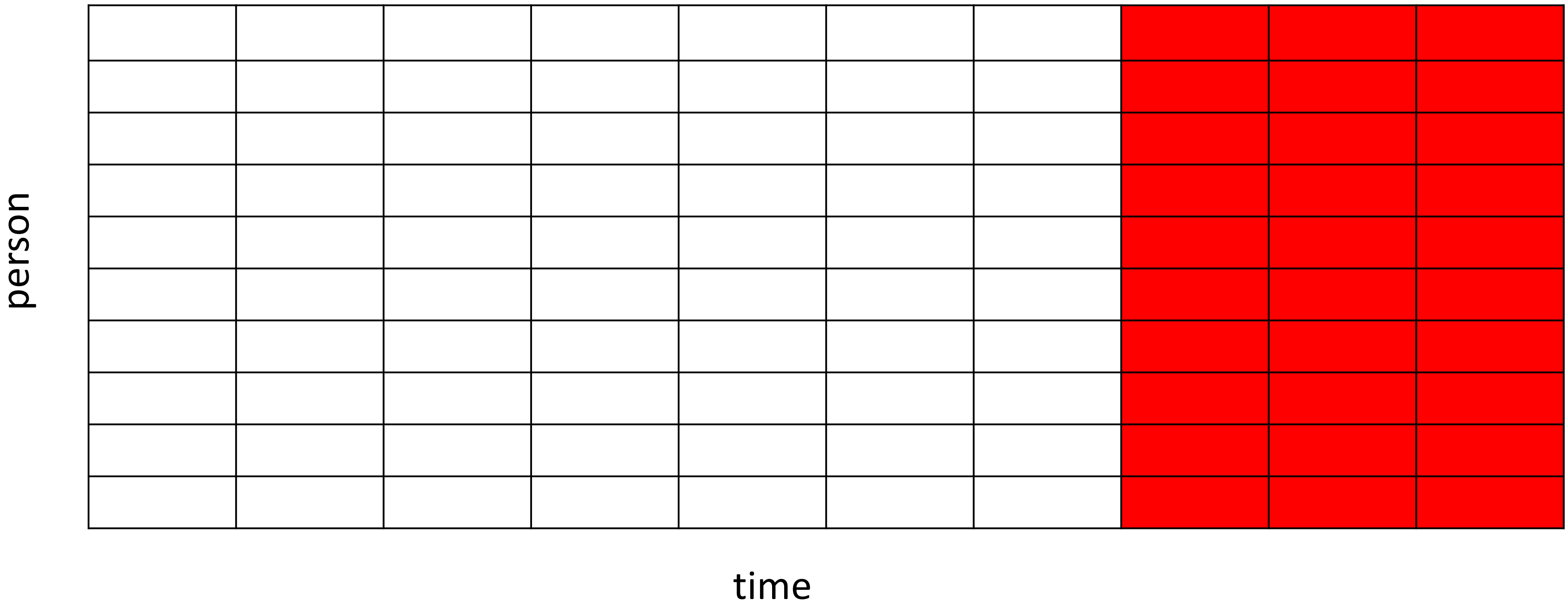


ASIDE: what is the ’correct’ test set for ESM?

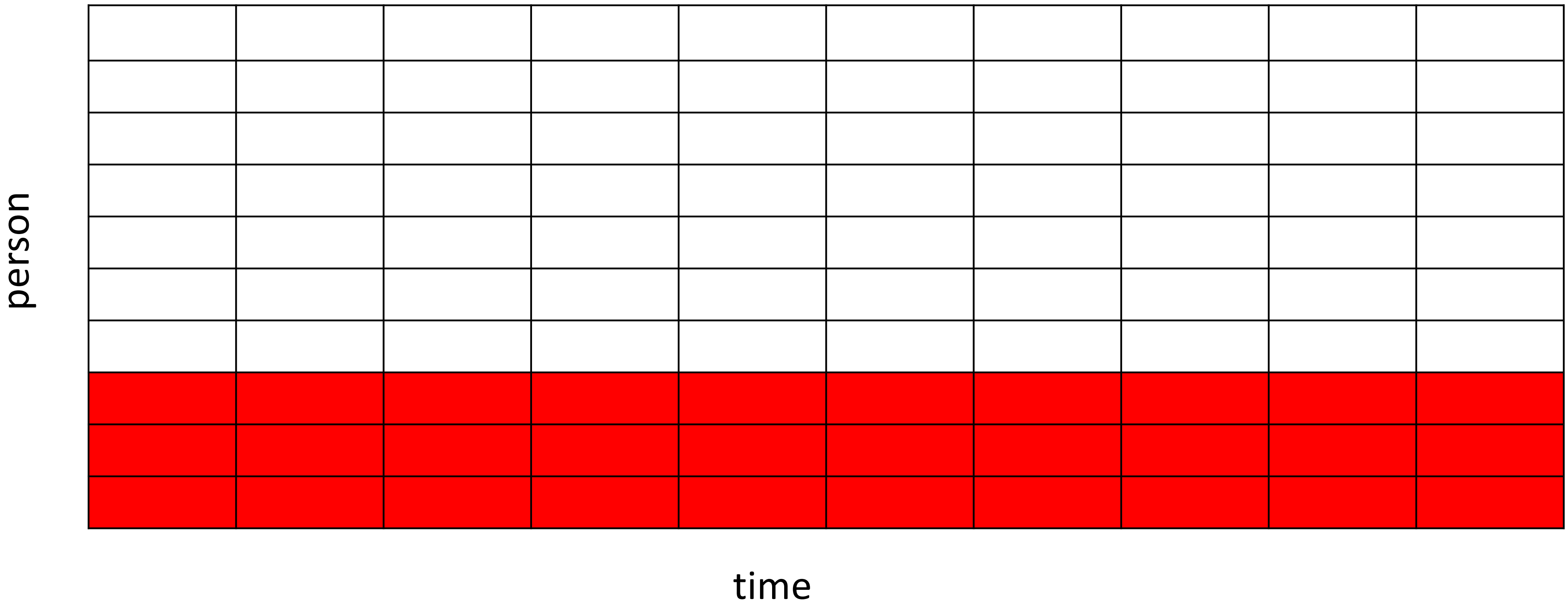


🟥									
		🟥							
			🟥						
								🟥	
						🟥			
									🟥
							🟥		
				🟥					
	🟥								
					🟥				

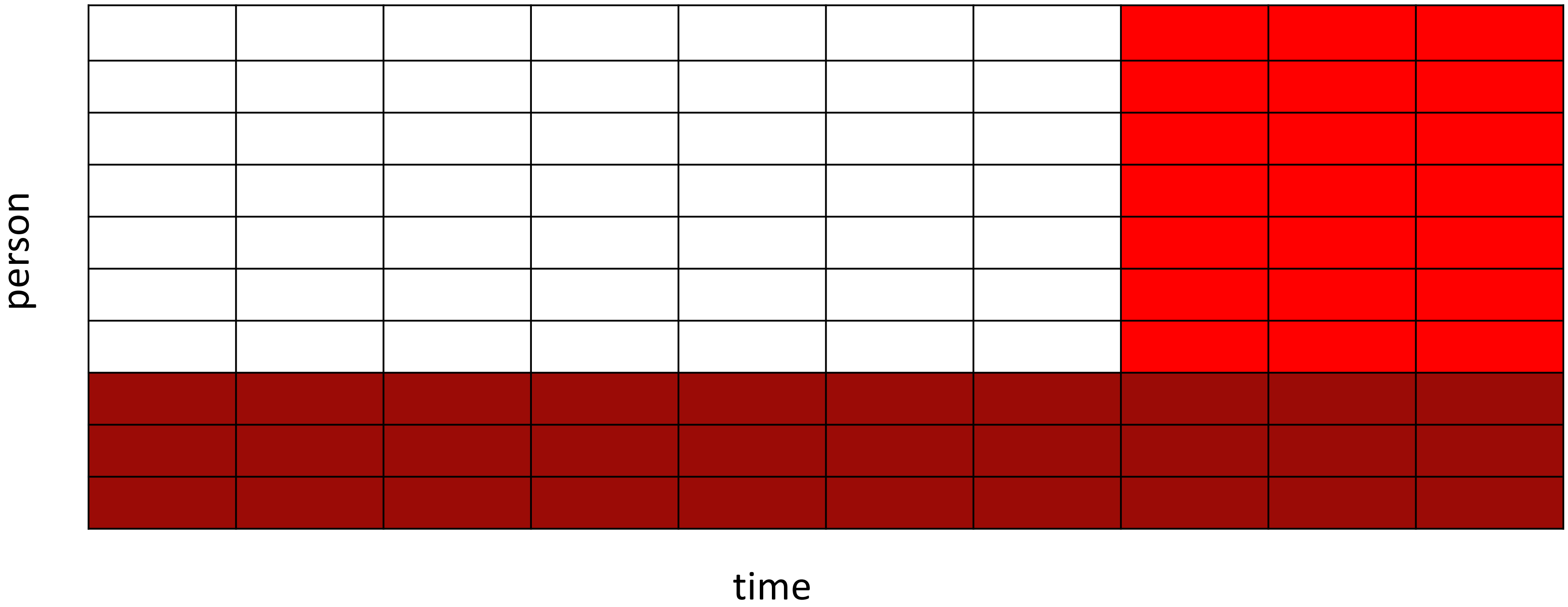
ASIDE: what is the ’correct’ test set for ESM?



ASIDE: what is the 'correct' test set for ESM?

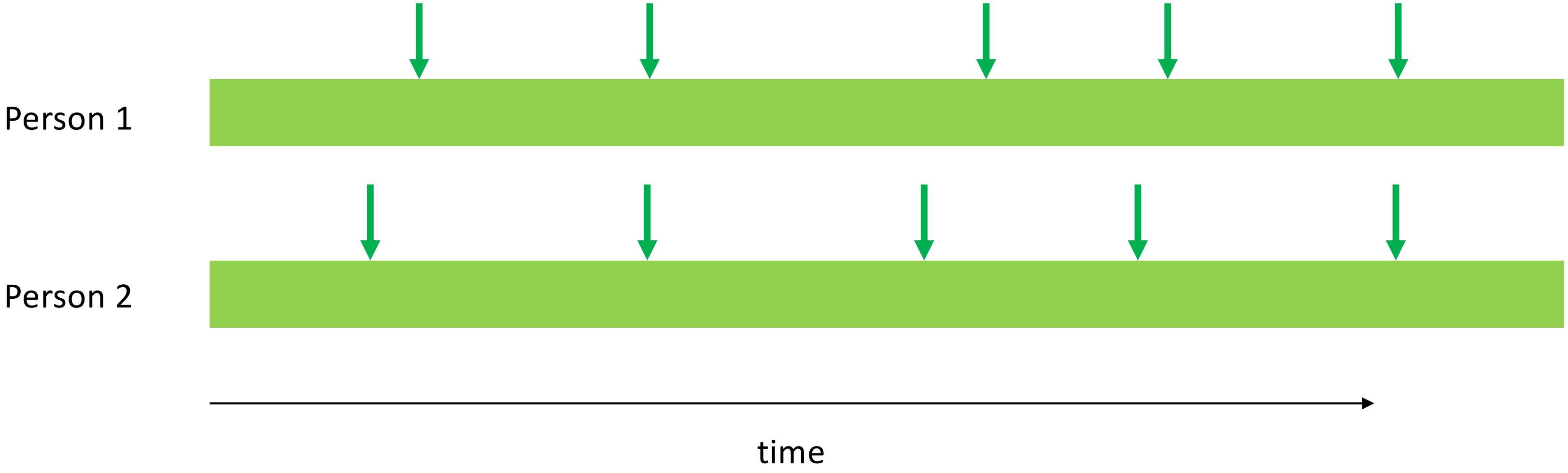


ASIDE: what is the ’correct’ test set for ESM?

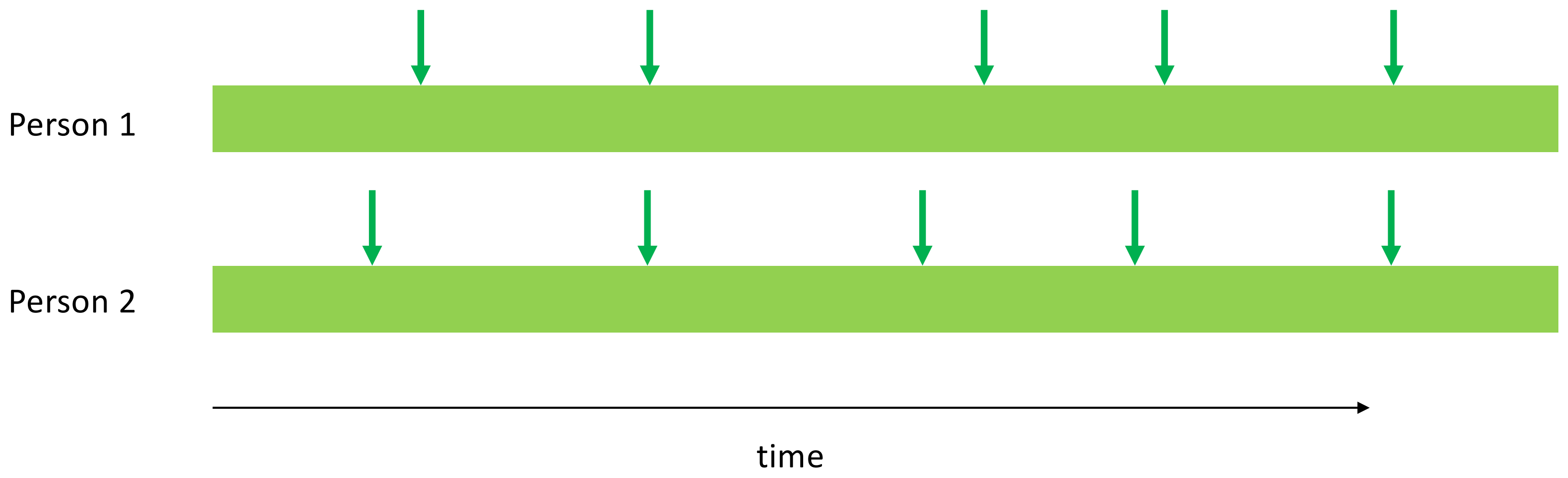


An Example: Digital Biomarkers of Stress

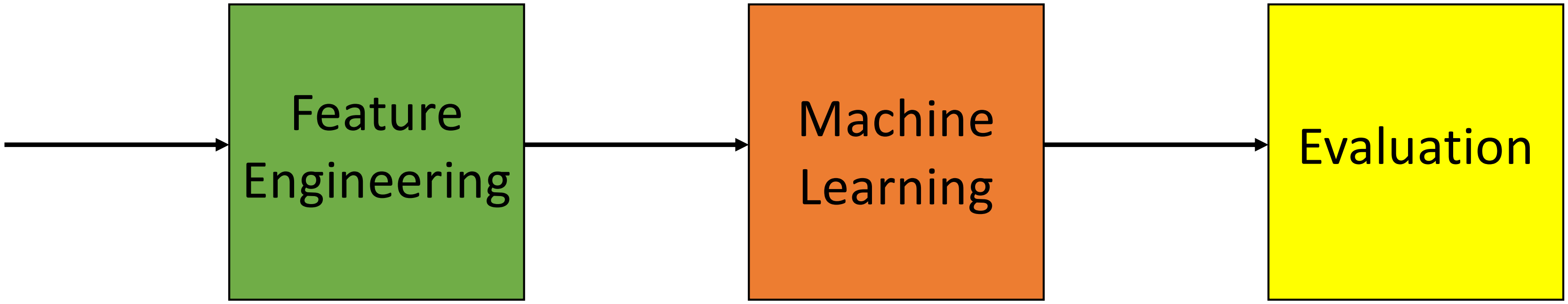
Digital Biomarkers of Stress: data structure



Q: How accurately can we predict momentary stress based on phone usage?



Our Data Science pipeline



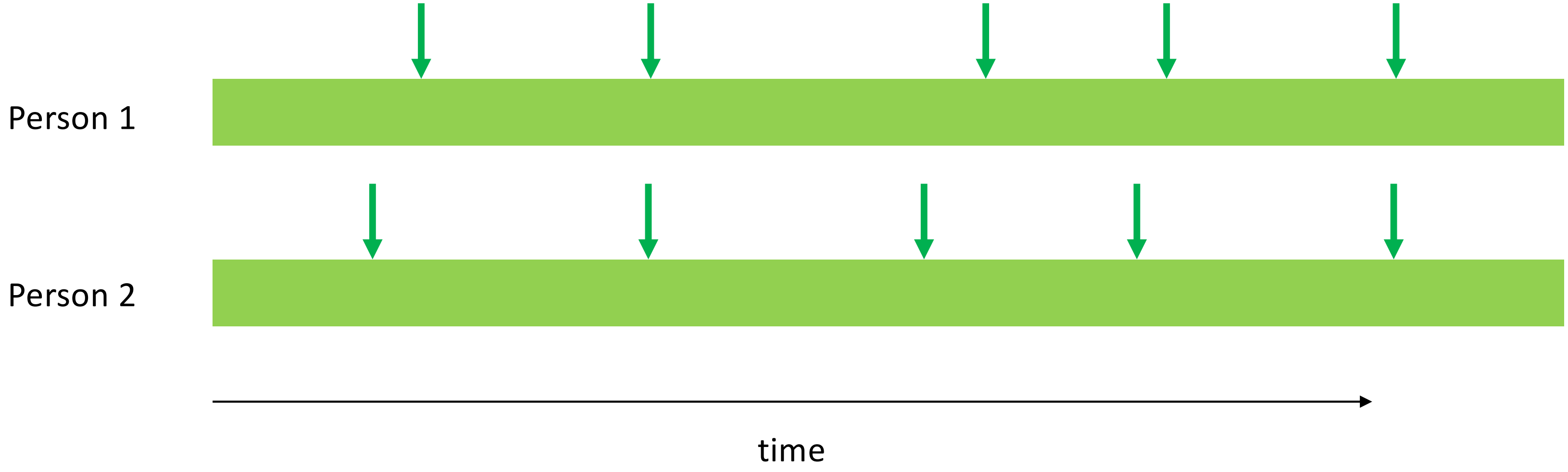
Feature Engineering

Machine Learning

Evaluation

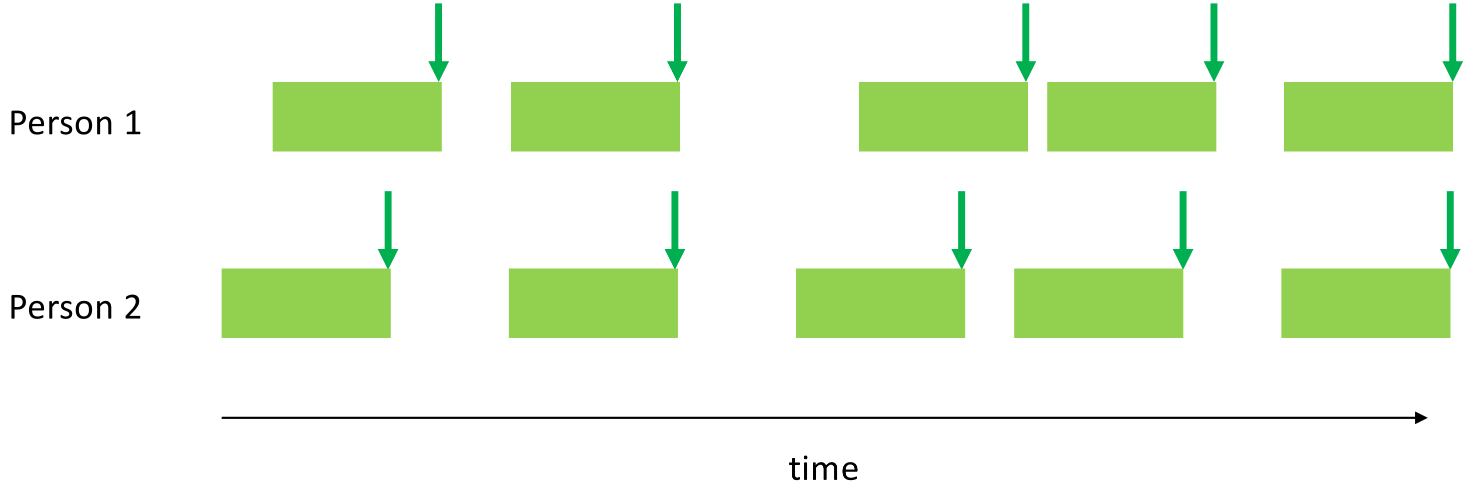
Engineering Features

Feature Engineering



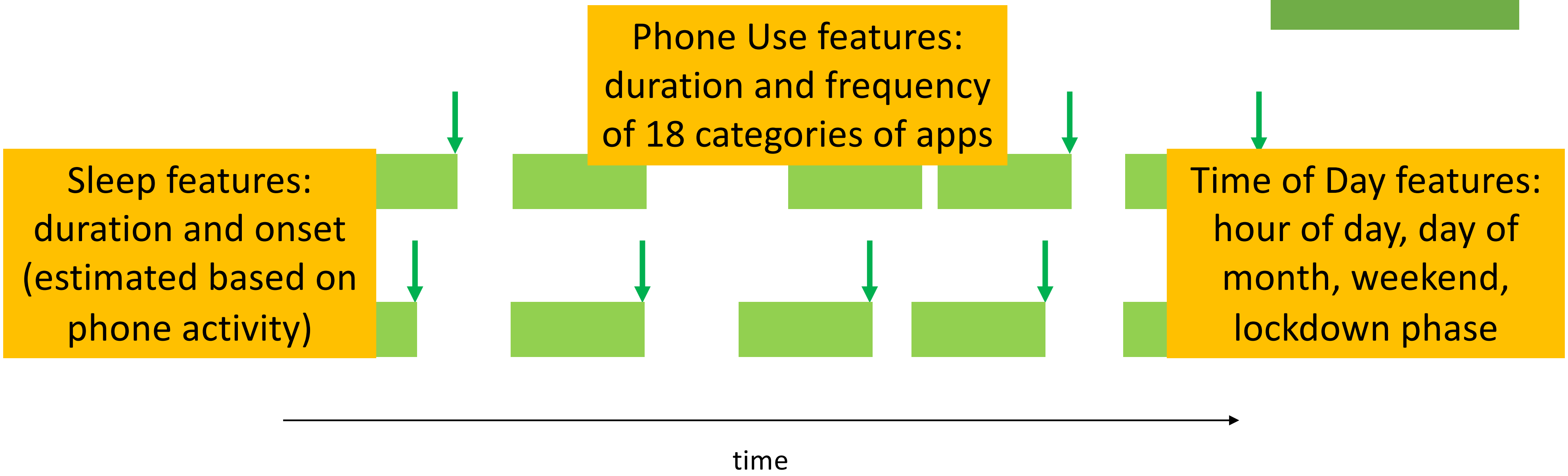
Engineering Features

Feature Engineering



Engineering Features

Feature Engineering

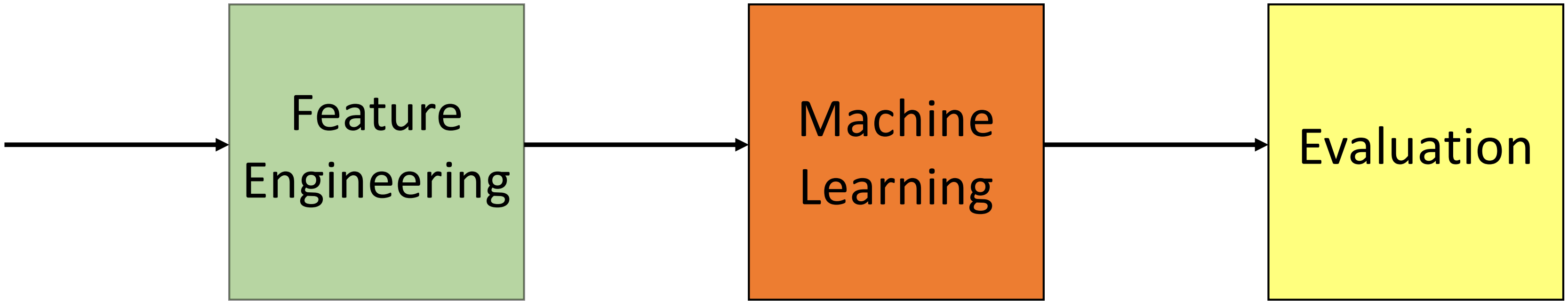


Phone Use features: duration and frequency of 18 categories of apps

Sleep features: duration and onset (estimated based on phone activity)

Time of Day features: hour of day, day of month, weekend, lockdown phase

Our Data Science pipeline



Feature Engineering

Machine Learning

Evaluation

Model: $\mathtt{y}_{\mathtt{train}} \sim \mathtt{x}_{\mathtt{train}}$

Lasso Regression

- Linear model
- Pressure to set many `beta` = 0
- Error based on RMSE

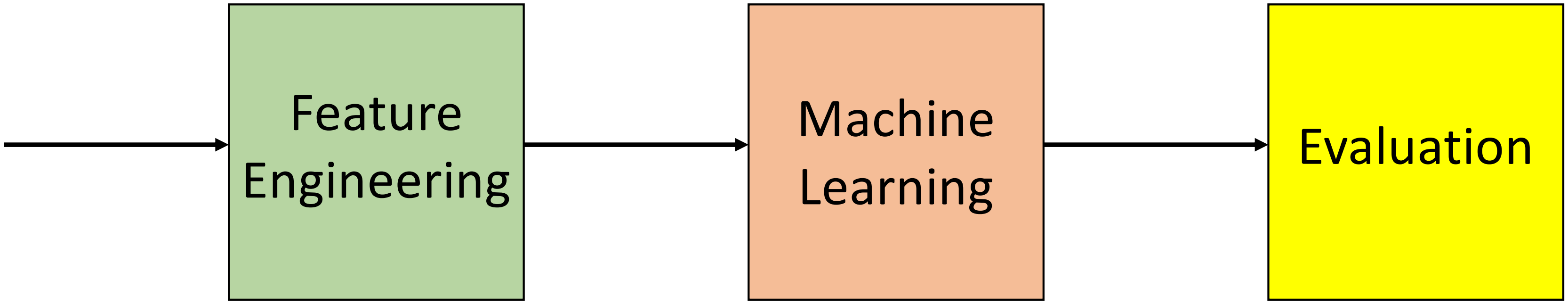
Random Forest

- Non-linear model
- Based on decision trees
- Train many trees, average the predictions (forest)
- Error based on R^2

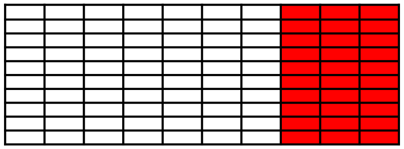
Support Vector

- Linear model
- Non-linear transformation of the input features
- Error not RMSE based

Our Data Science pipeline



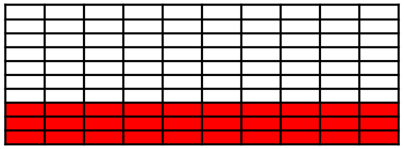
Ideographic Models



80%, 20%

Q: Can we predict this person in the future?

Nomothetic Models



5 folds of 45 people

Q: Can we predict for a new person?

$$\mathtt{ABS}(y_{test},\ \mathtt{Model}(X_{test}))$$

$$\mathtt{Spearman\ rho}(y_{test},\ \mathtt{Model}(X_{test}))$$

Results

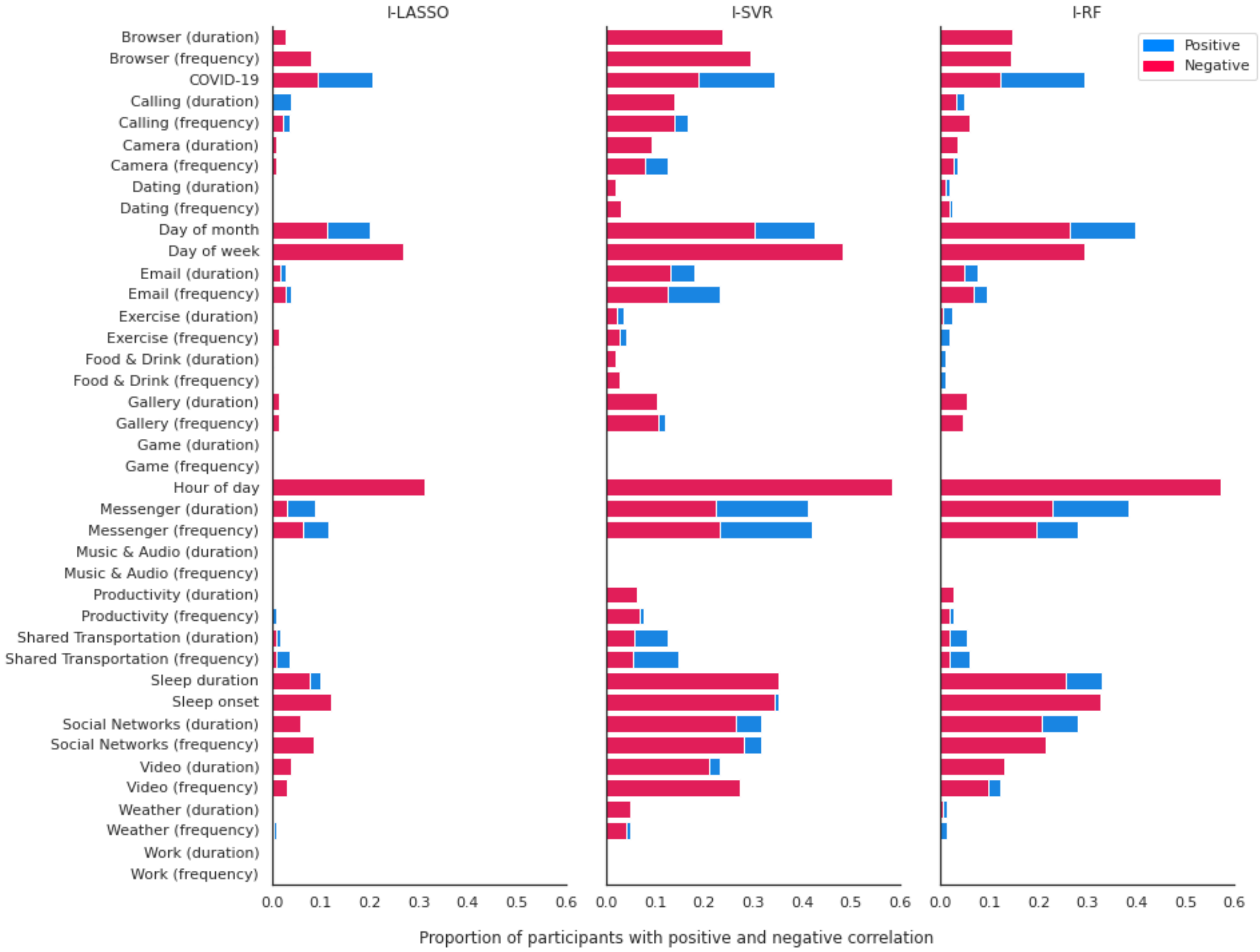
Ideographic (predicting future stress for a person)

- Correlation metric:
  - Random Forest: median rho = 0.10, 20.5% people rho significantly > 0
- Absolute error metric:
  - Support Vector: median error = 0.85, best model for 38% of people

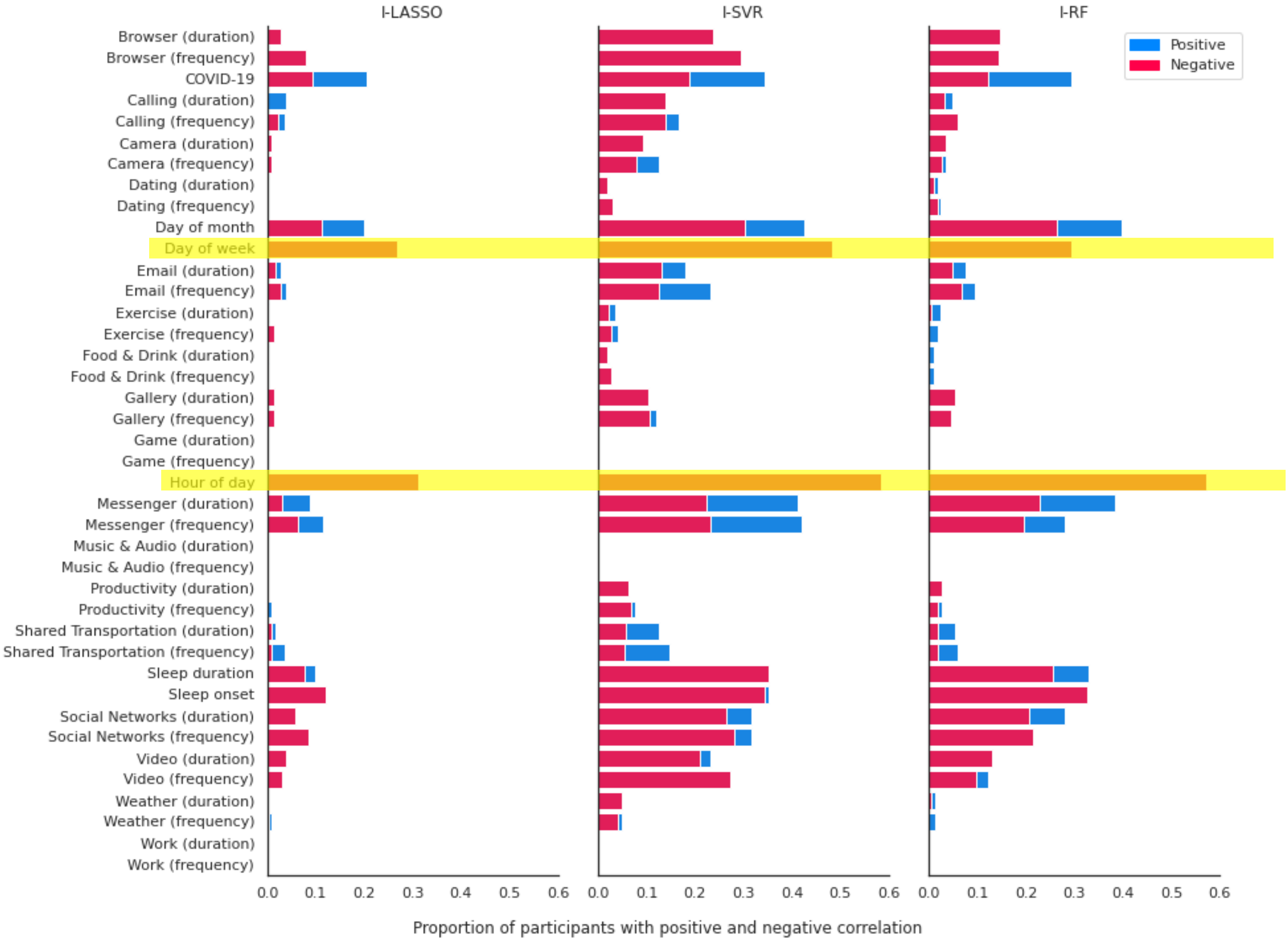
Nomothetic (predicting for new, unseen people)

- Correlation metric:
  - Random Forest: median rho = 0.18, 55.8% people rho significantly > 0
- Absolute error metric:
  - Baseline model: median error = 0.83, best model for 89% of people

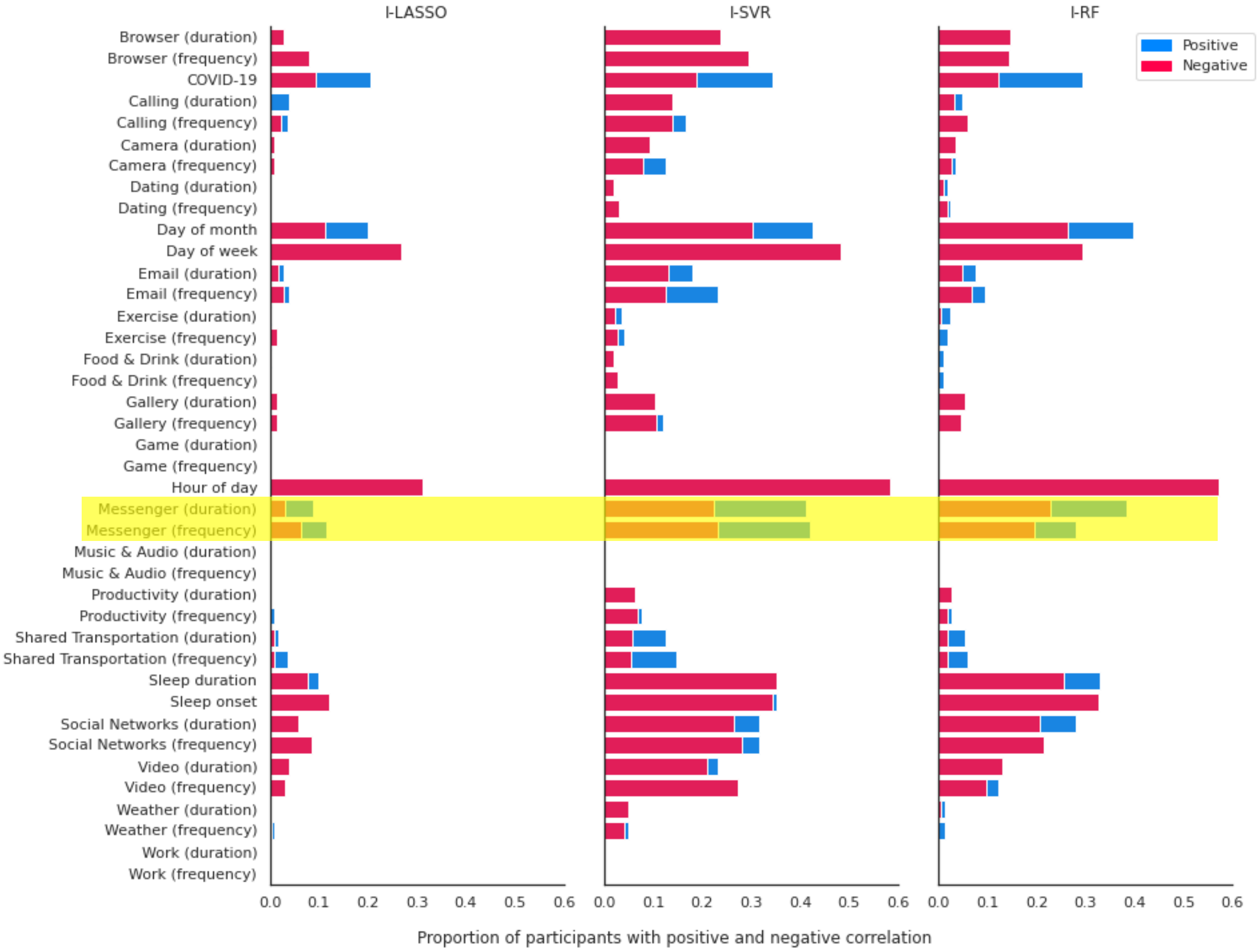
Results: Individual Differences



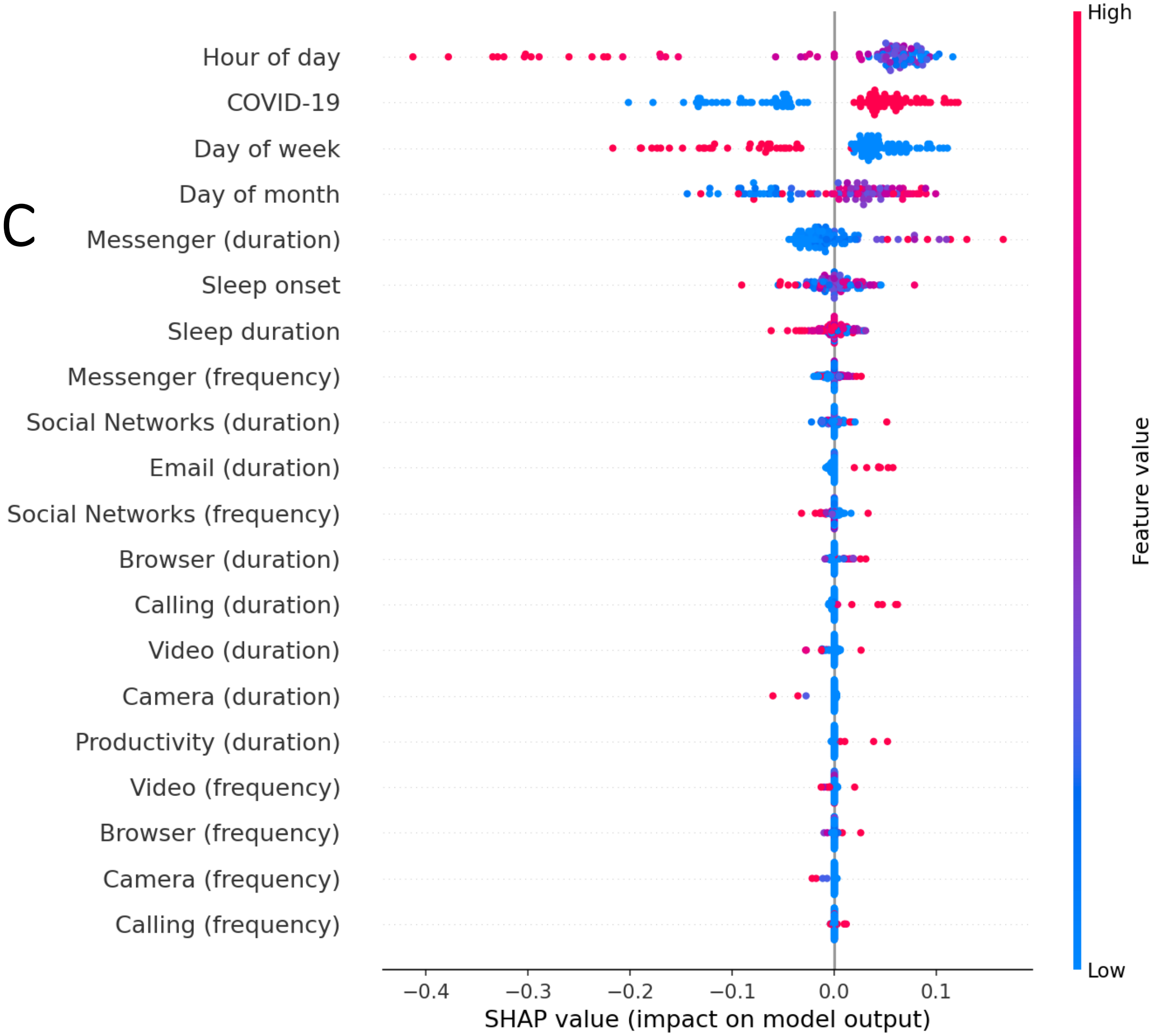
Results: Individual Differences



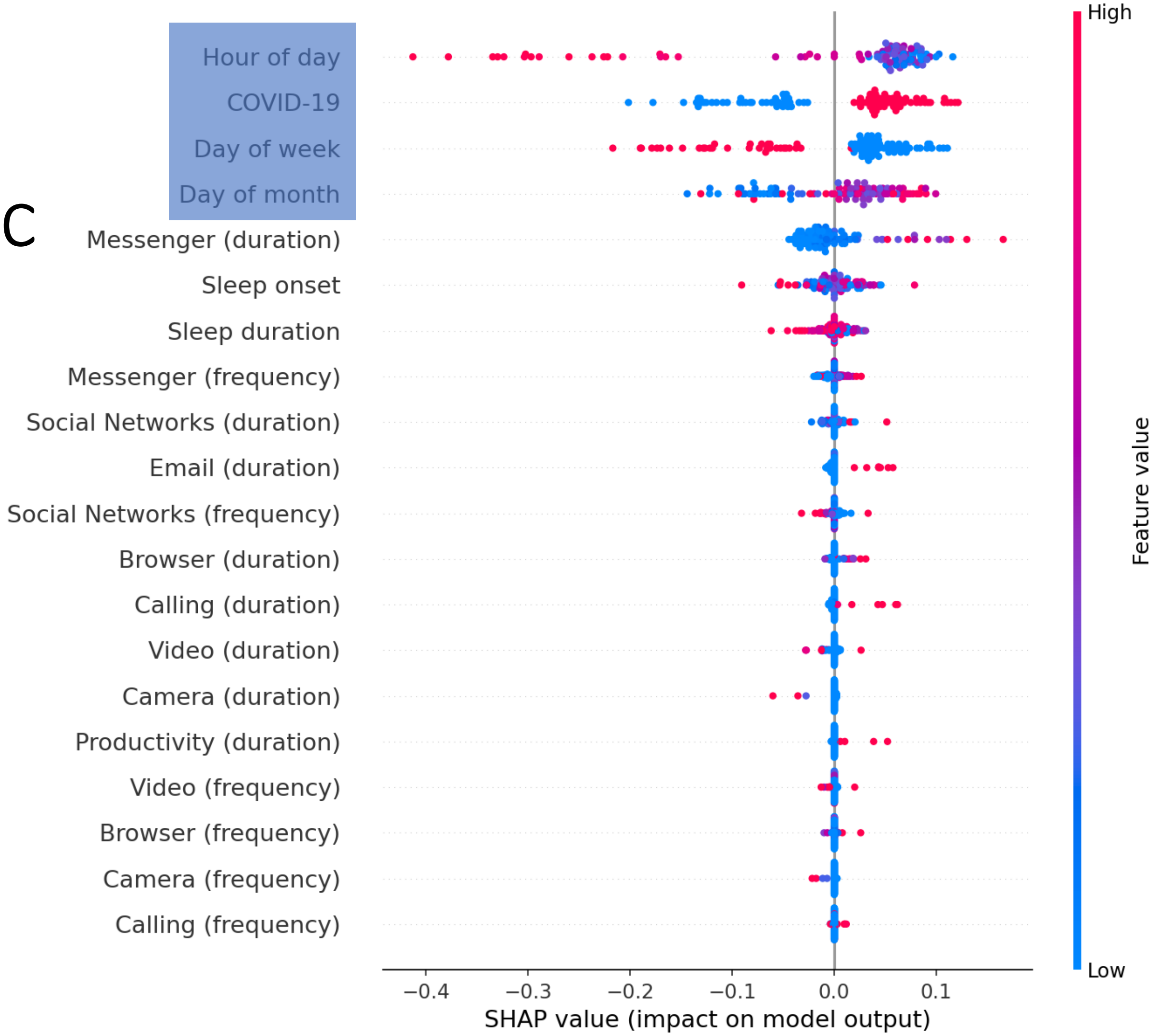
Results: Individual Differences



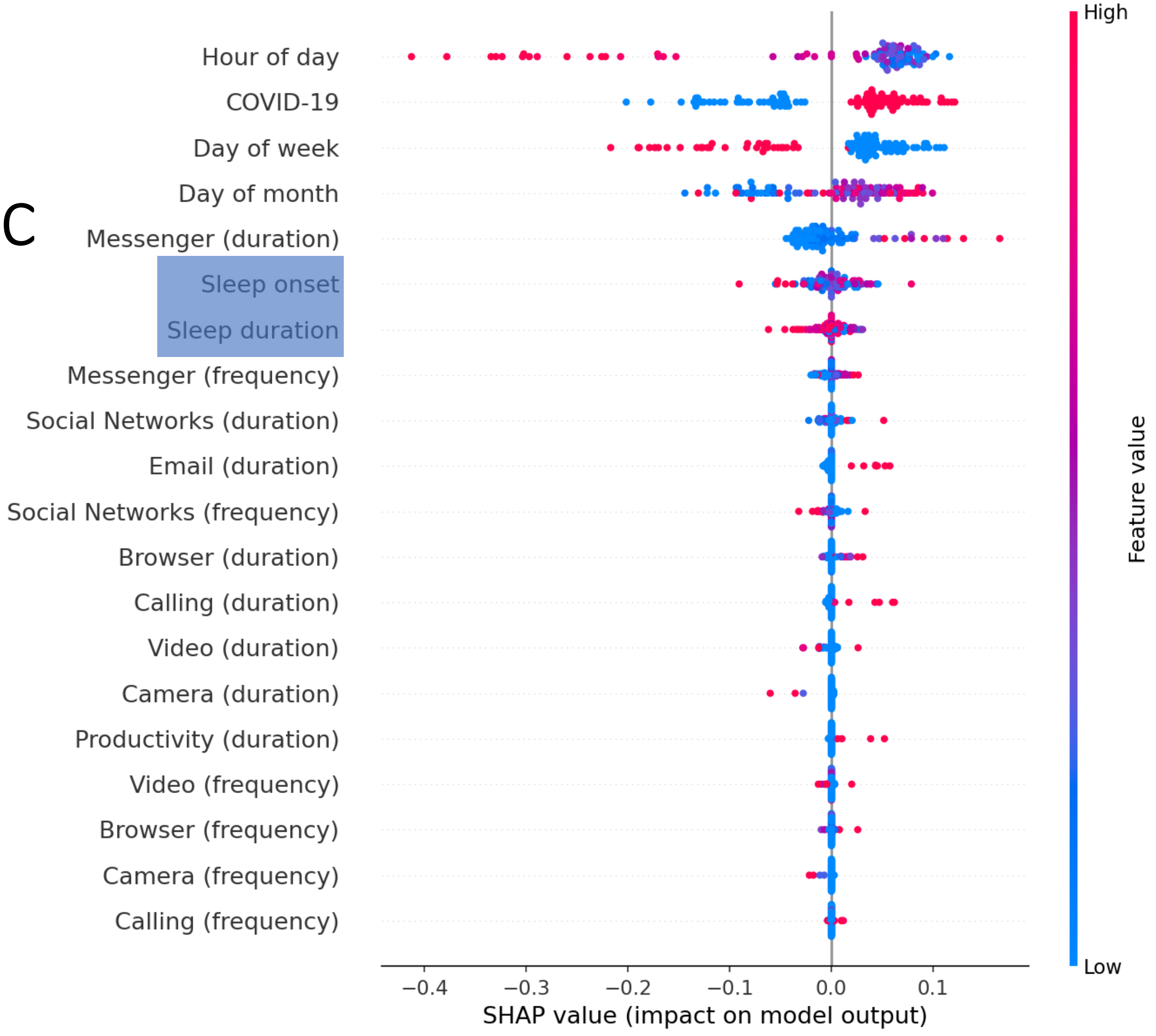
Results: Nomothetic



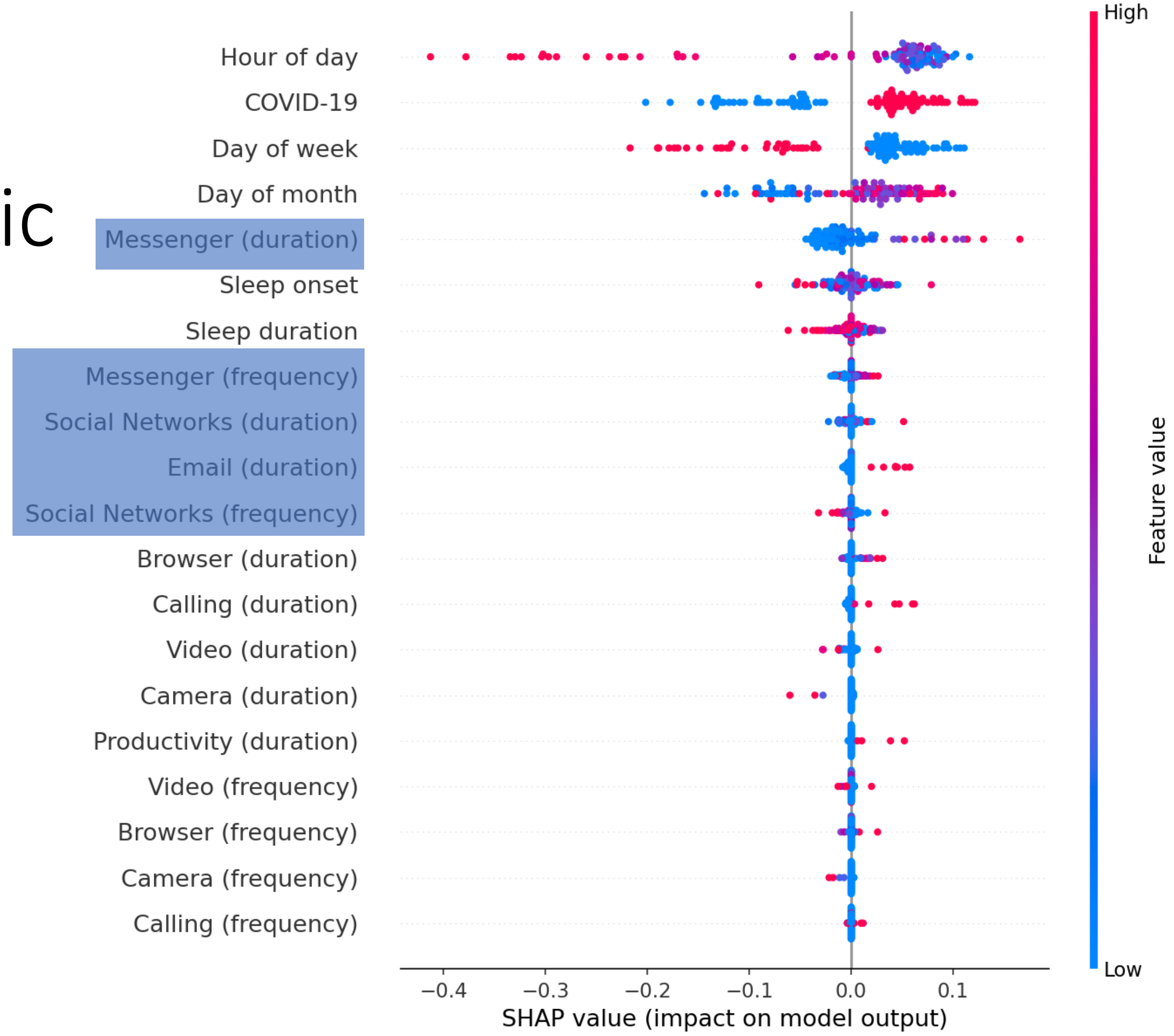
Results: Nomothetic



Results: Nomothetic



Results: Nomothetic



Where to next?

- Phone application data alone can be limited in utility and scope
    - Combinations of data streams (ESM, phone, sensors) can provide a more rich digital footprint
- New research group in CSAI: AI & Data Science for Health & Well-being
    - Expertise in Sequential Pattern Mining and Machine Learning / Deep Learning

Thanks!

TILBURG UNIVERSITY

Understanding Society

IMPACT Program:
Health and Well-being

George Aalbers
Tilburg University

Mariek vanden Abeele
Ghent University

Loes Keijsers
Erasmus University
Rotterdam

Web: drewhendrickson.github.io
LinkedIn: drew-thomas-hendrickson