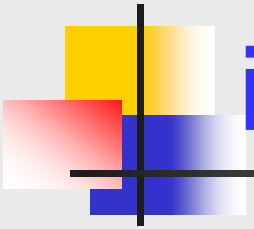
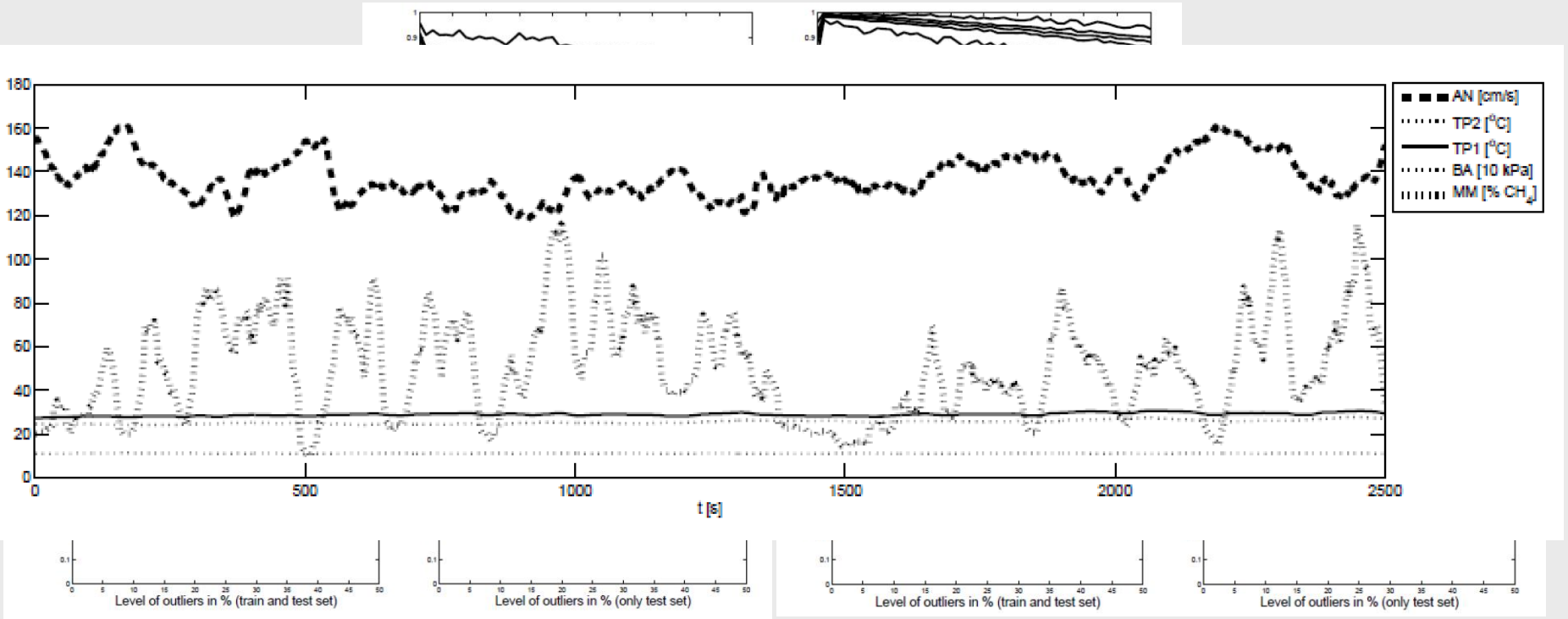
ETL2 (wartości brakujące i odstające)



- Od jakiej liczby (%) wartości odstających (brakujących) zaczyna się psuć model diagnostyczny?
- Które z znanych metod dobrze identyfikują outliery?
- Czy intuicyjne metody zastępowania wartości odstających i brakujących są naprawdę gorsze od metod zaawansowanych obliczeniowo?
- Badania: wszystkie dane, przyrostowo, w przesuwającym się oknie czasowym
- **Problem otwarty**: identyfikacja wartości odstających w wielowymiarowym strumieniu danych

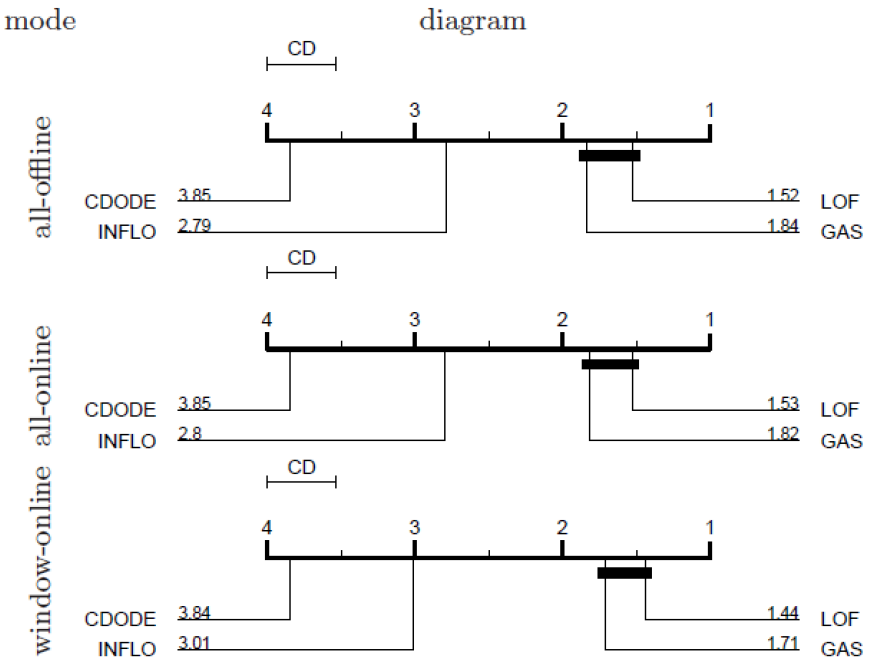
ETL2 (wartości brakujące i odstające)



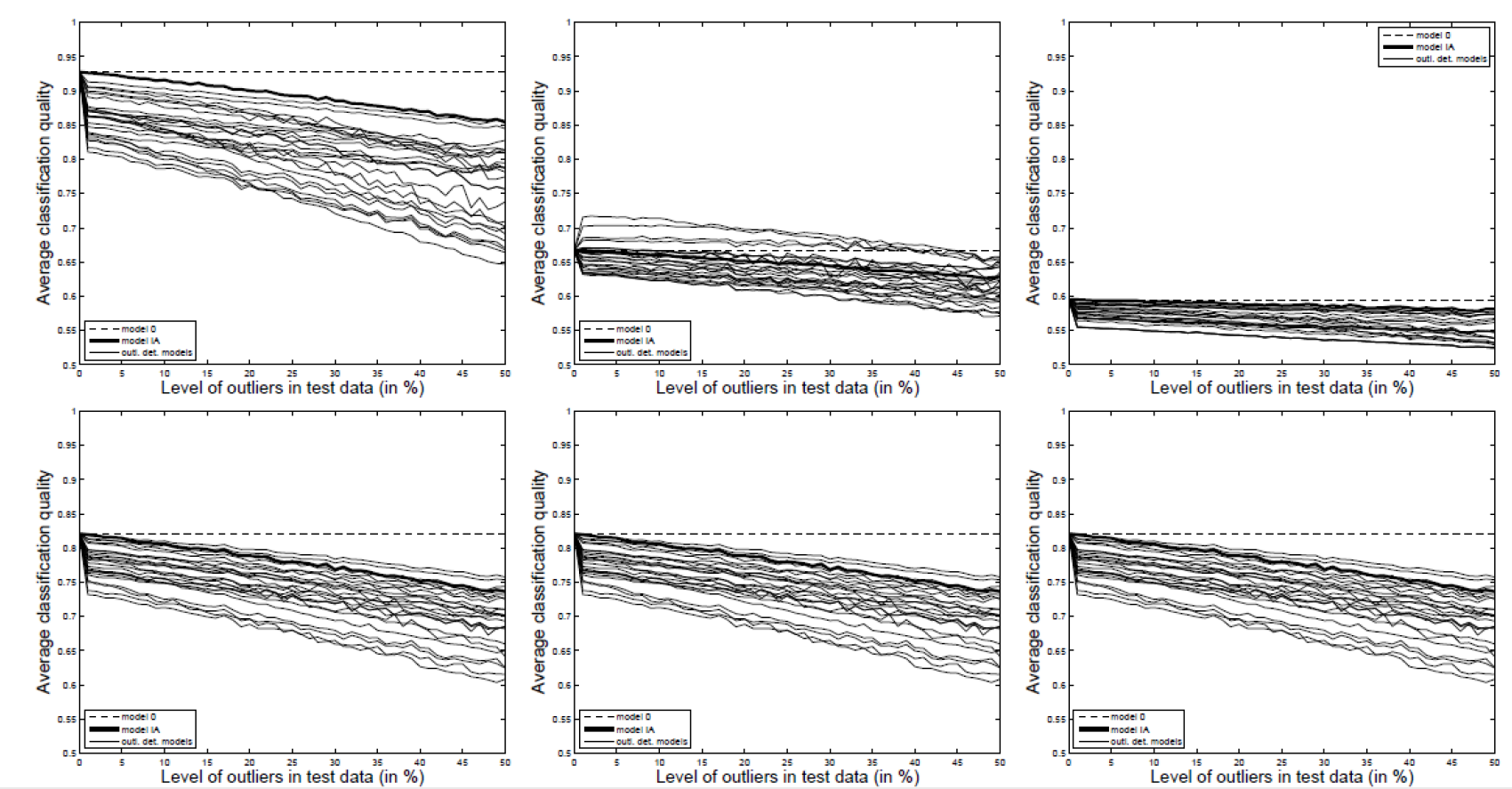
ETL2 (wartości brakujące i odstające)

Method	an average AUC		
	all-offline	all-online	window-online
CDODE	0.719	0.734	0.716
GAS	0.848	0.831	0.824
INFLO	0.781	0.803	0.795
LOF	0.855	0.839	0.832

Method	all-offline vs. all-online	
	wins/losses	Wilcoxon p-value
CDODE	38/62	0.011726
GAS	74/26	0.000178
INFLO	37/63	0.000121
LOF	69/31	0.001102

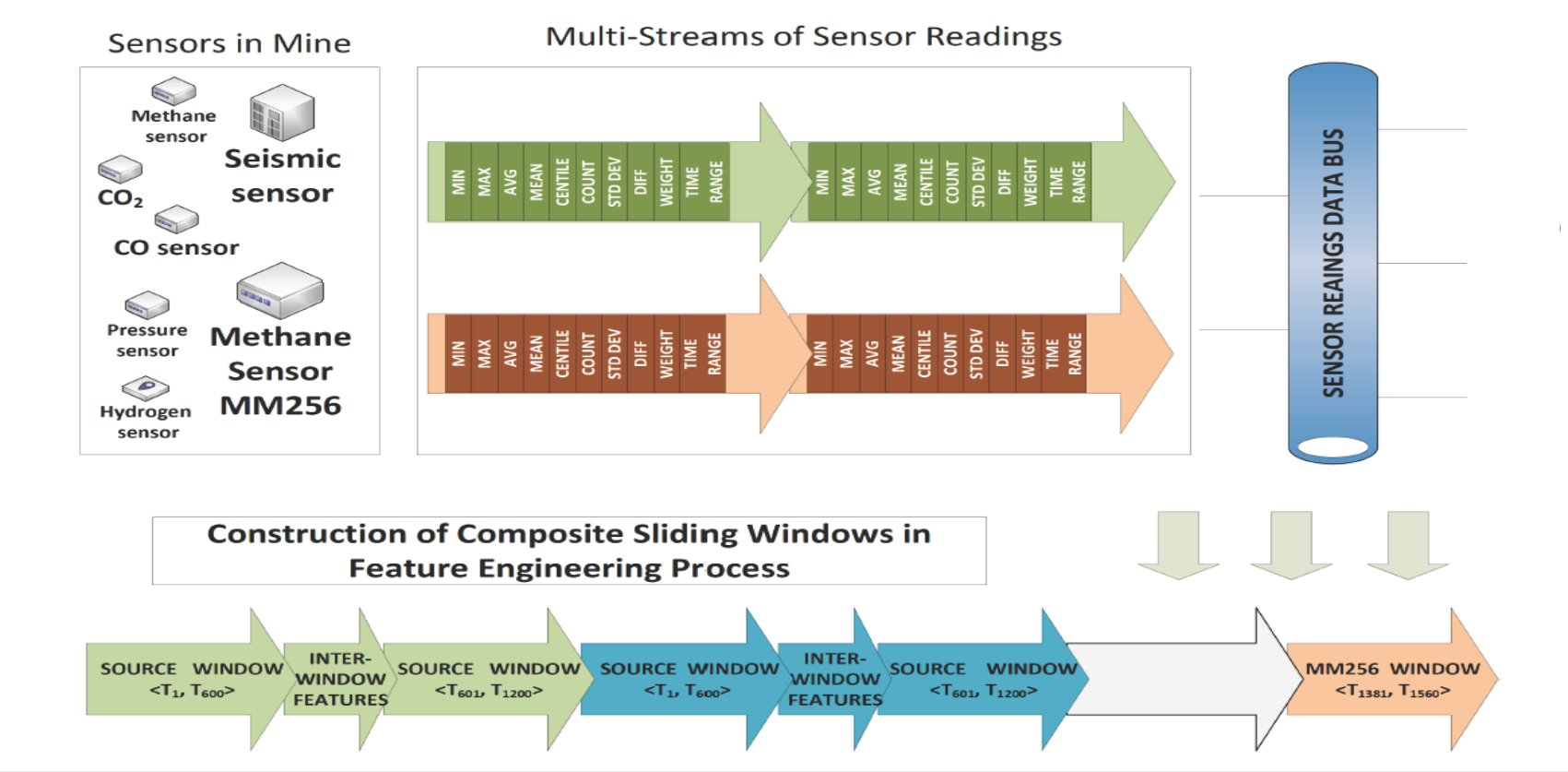


ETL2 (wartości brakujące i odstające)



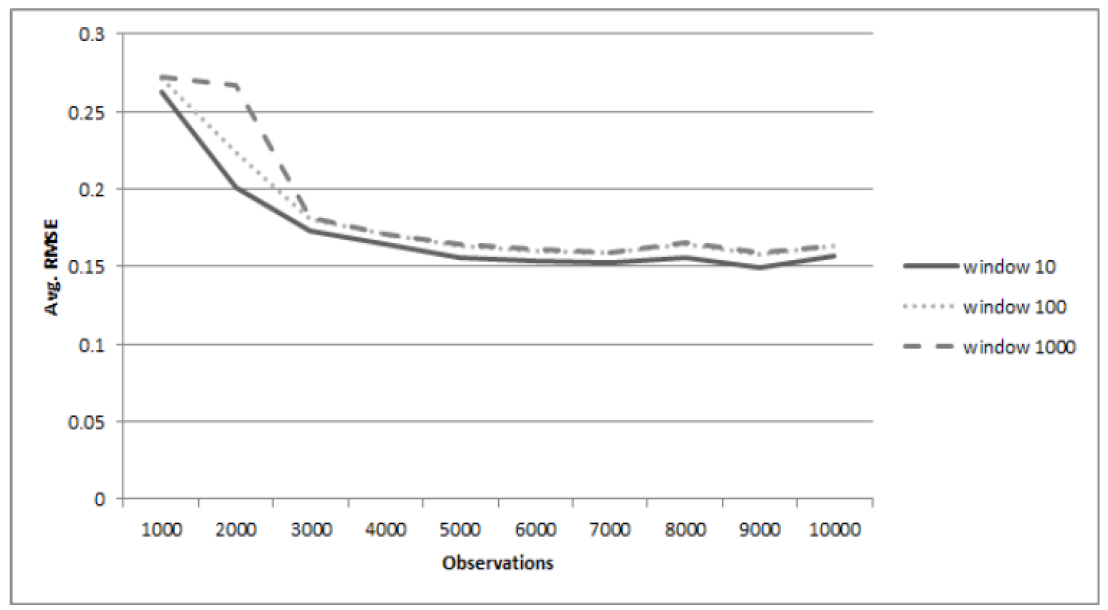
ETL2 (ekstrakcja i selekcja cech)

- Ekstrakcja - zamiana szeregu czasowego na zagregowane atrybuty



ETL2 (ekstrakcja i selekcja cech)

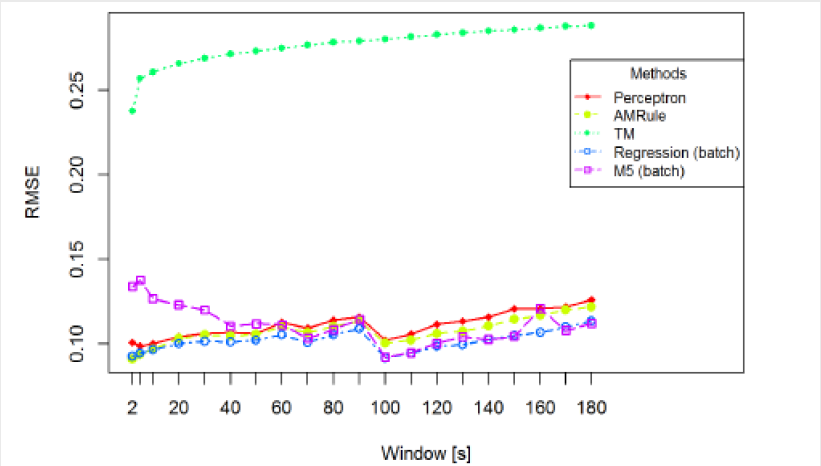
- Dla klasyfikacji to się sprawdza (patrz konkursy)
- Dla regresji:



- **Kwestia otwarta**: Czy tak jest w przypadku innych danych? Czy horyzont prognozy ma na to wpływ?

ETL2 (ekstrakcja i selekcja cech)

- Dla regresji:



(a) *whole* approach