INNOVATION INSIGHT
Health AI with Constrained Data Collection: Technologies and Applications

Jiangtao Wang
Associate Professor of Health Data Science
Research Centre for Intelligent Healthcare
Coventry University

Health AI with Constrained Data Collection: Technologies and Applications

Dr Jiangtao Wang , Associate Professor,

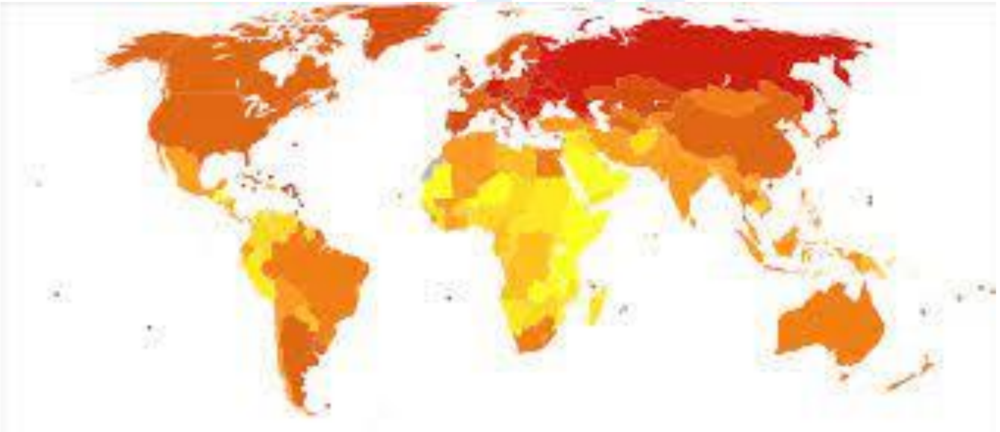
Centre for Intelligent Healthcare, Coventry University

@ Intelligent Health, Basel, Switzerland

Continuous health monitoring at both individual and population level is a key research problem in digital/intelligent health

Individual health monitoring

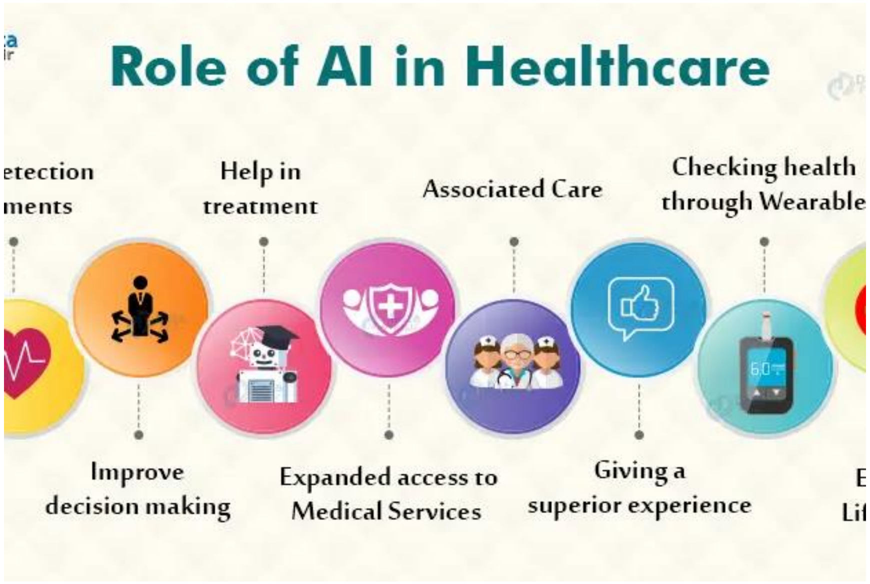
Population health monitoring



Two enablers: UbiComp + AI

UbiComp for healthcare data collection (sensing)

AI for health/healthcare data analytics & prediction



One key challenge in the pathway

Ideal: huge data, good annotation → powerful ML models to achieve satisfactory performance

- Data accumulation in EHR, mobile, wearable, etc.

Reality: limited data, most the data are **unlabelled** → fail to meet the application requirement

- Collecting large amount of labelled data and build the model from the scratch: **expensive and time-consuming**

Coventry University

high cost

limited cohort size

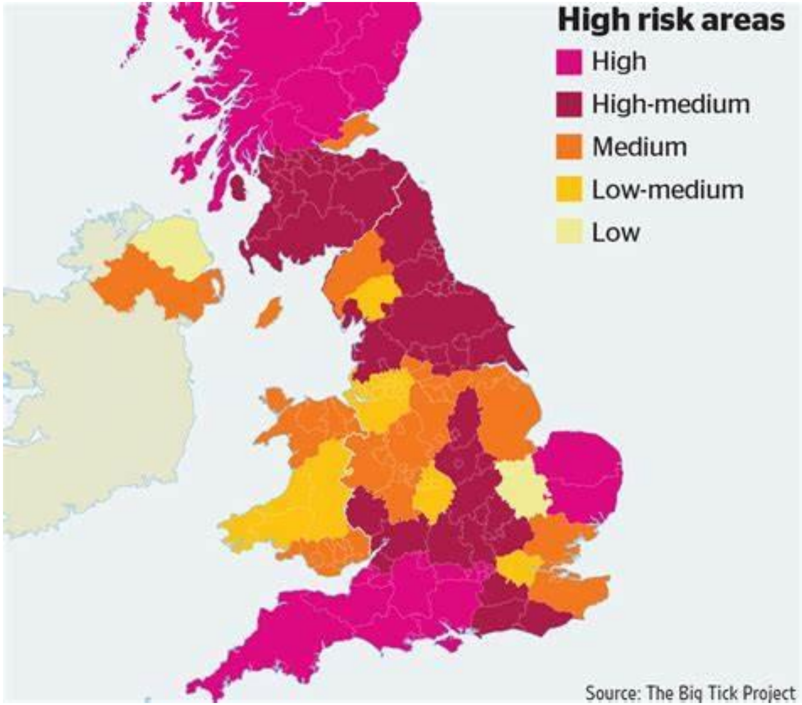
uncertain participation willingness

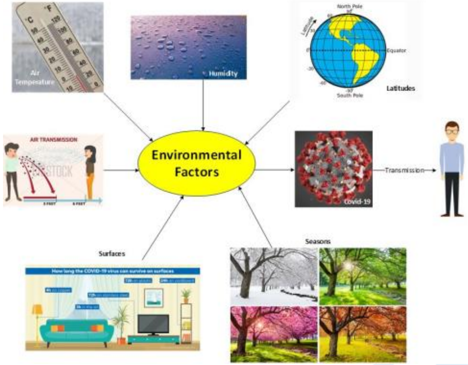
law & policy constraint

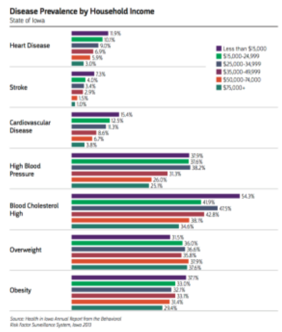
Factors leading to this challenge

How to build an intelligent healthcare system with minimal data collection?

Area-level population health profiling (a real-world problem from NHS)





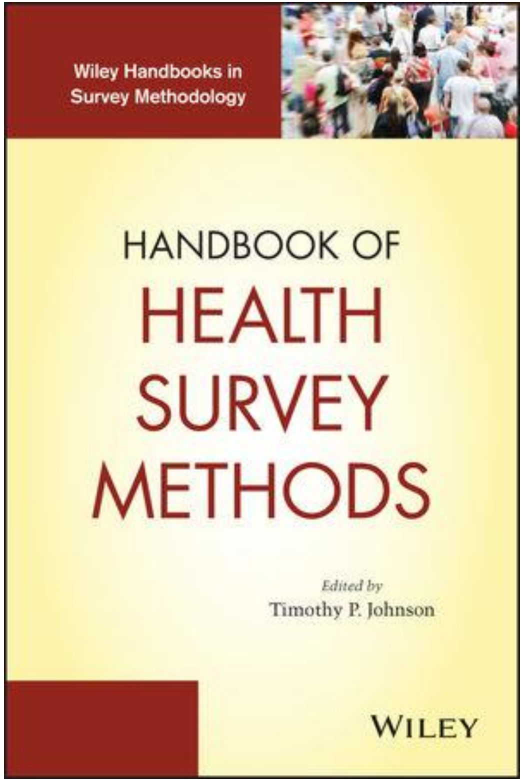


teaching and learning
Teaching Excellence Framework (TEF)

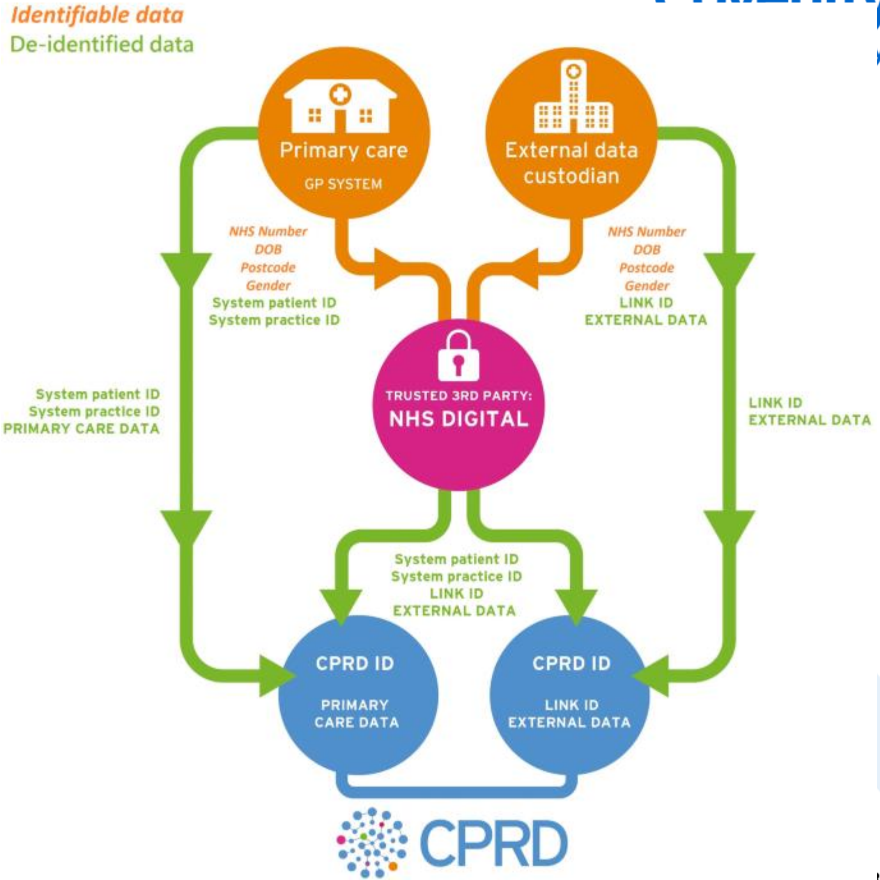
Teaching Quality
The Times and Sunday Times Good University Guide 2017

d No.12
UK University
Guardian University Guide 2018

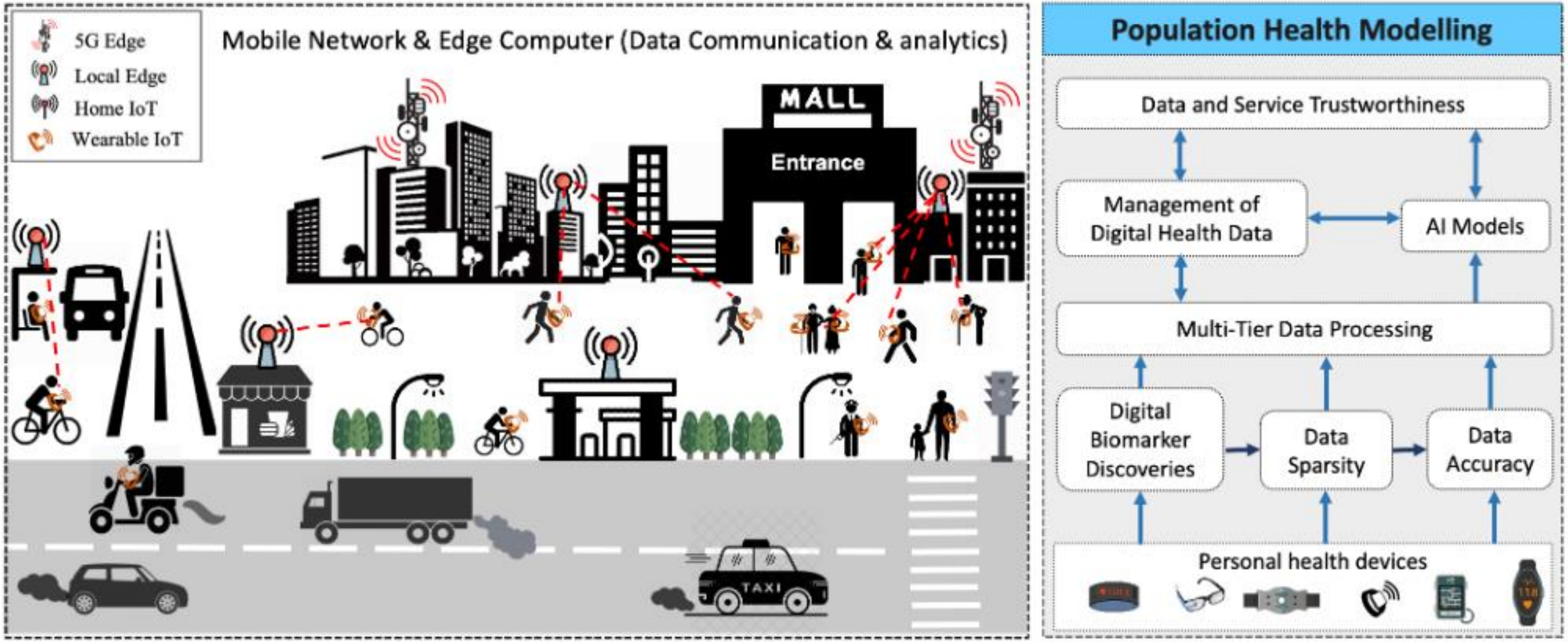
Health Surveys



Population health data linkage and integration



Pervasive & Mobile Computing

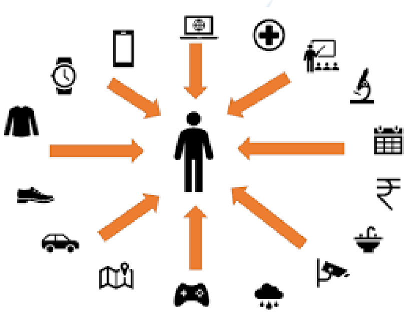


Summary of limitations

health surveys

clinic data integration

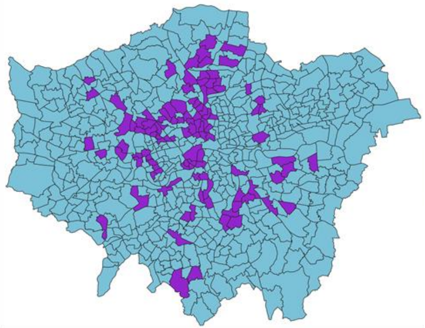
Mobile/pervasive Comp

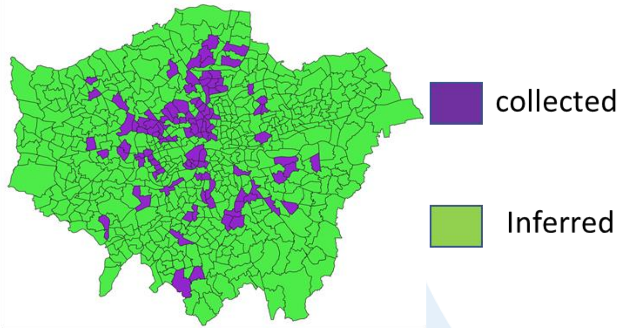


<u>*No matter what approach you have adopted*</u>

- *limited spatial coverage*: ***unknown/not usable (not accurate)***
- *Unknown areas:* ***hard-to-reach population, health inequality***

Compressed Population Health (CPH): basic idea





Given a target region for health profiling

CPH can select a subset of grids (where stakeholders will do traditional profiling)

CPH Infers the profiles in un-selected grids

Ongoing project in my team **supported by EPSRC New Investigator Award**

Two types of data correlations

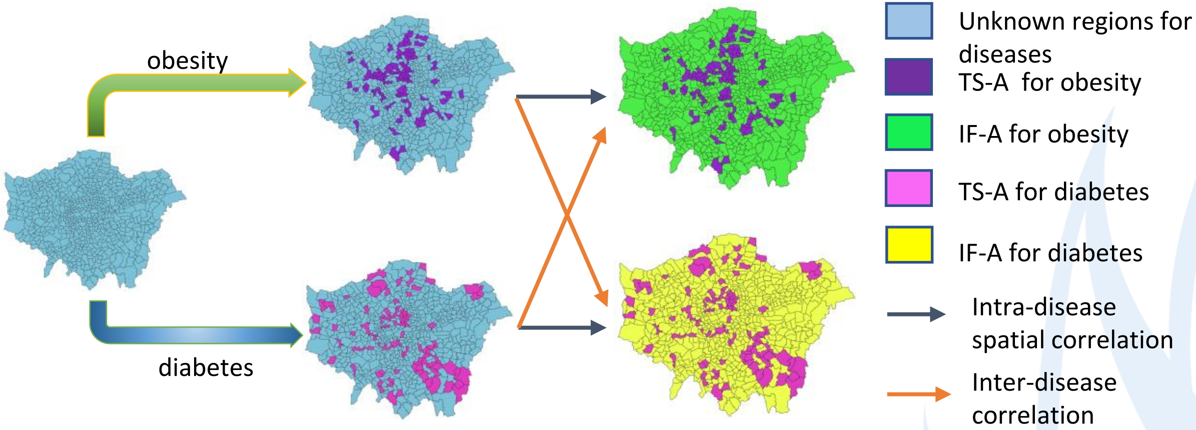
(a) Intra-Disease Spatial Correlations

- a number of studies have highlighted the role of ***neighbourhood effects*** on health
- **near regions are more similar in some health indicators than the distant ones**

(b) Inter-Disease Correlations

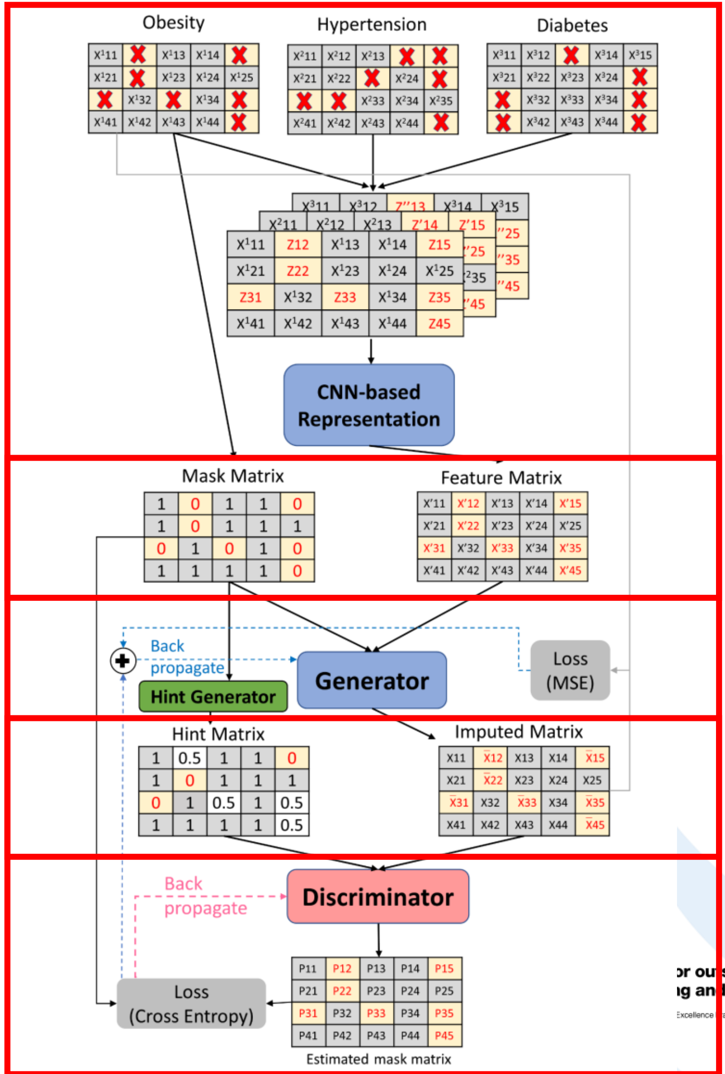
- ***Multimorbidity***, commonly defined as the ***co-presence of two or more chronic*** *conditions*
- **statistics for different types of disease may also correlate with each other.**
  - e.g., regions with higher obesity rate are more likely to have higher rates of heart disease and cancers.

Jointly use intra- and inter- disease correlations



CNN-based representation learning (extracting two types of correlations)

Generative Adversarial Network (GAN) for data reconstruction



Coventry University

- ***GAN: two neural networks*** contest with each other
  - ***Generator***: learns to generate new data with the same statistics as the training set
  - ***Discriminator***: another neural network that is able to tell how much an input is "realistic",

Top 4 for Student Experience Teaching Quality
The Times and Sunday Times Good University Guide 2017

Ranked No.12 UK University
Guardian University Guide 2018

Datasets

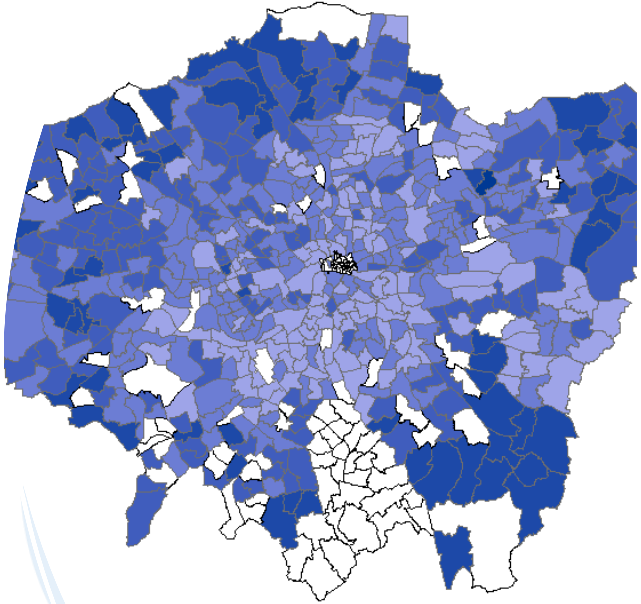
Coventry University

Dataset of Ward Boundaries of London

- The dataset includes names, shapes and codes of **630 grids (wards)** in London.

Chronic Diseases Prevalence Dataset

- It contains prevalence rate of **17 chronic diseases**: **from 2008 to 2017** of London ward level.



Results for missing data completion

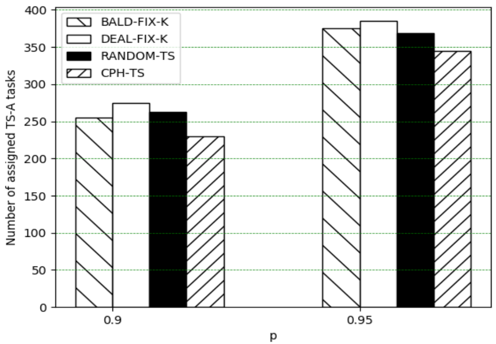
- Our CNN+GAN model outperforms all baseline ones across all disease in all evaluation metrics and settings (e.g. data missing rate).

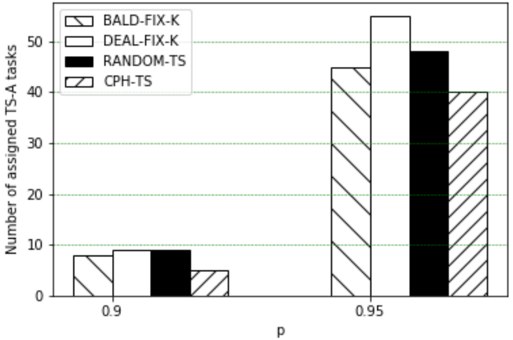
Methods	2016						2017					
	$\mathcal{R}$ = 0.1		$\mathcal{R}$ = 0.3		$\mathcal{R}$ = 0.5		$\mathcal{R}$ = 0.1		$\mathcal{R}$ = 0.3		$\mathcal{R}$ = 0.5	
	RMSE	MAE	RMSE	MAE	RMSE	MAE	RMSE	MAE	RMSE	MAE	RMSE	MAE
CF	0.1661	0.1345	0.1657	0.1340	0.1640	0.1298	0.1983	0.1625	0.2010	0.1659	0.2028	0.1702
Average(spatial)	0.1478	0.1202	0.1400	0.1153	0.1401	0.1149	0.1581	0.1310	0.1525	0.1288	0.1444	0.1229
Median(spatial)	0.1518	0.1252	0.1367	0.1110	0.1355	0.1100	0.1509	0.1233	0.1435	0.1201	0.1370	0.1150
NMF	0.1518	0.1180	0.1346	0.1064	0.1412	0.1113	0.1661	0.1331	0.1513	0.1208	0.1330	0.1054
TD	0.1403	0.1045	0.1275	0.1014	0.1250	0.1002	0.1304	0.0997	0.1221	0.0970	0.1181	0.0923
Linear Regression	0.1026	0.0763	0.0947	0.0730	0.0927	0.0687	0.1132	0.0934	0.0887	0.0714	0.0853	0.0671
Auto-encoder	0.0857	0.0616	0.0817	0.0597	0.0821	0.0597	0.0772	0.0575	0.0681	0.0520	0.0654	0.0496
stKNN	0.0794	0.0557	0.0752	0.0546	0.0732	0.0528	0.0739	0.0520	0.0632	0.0472	0.0609	0.0459
Median(temporal)	0.0830	0.0564	0.0769	0.0537	0.0760	0.0525	0.0776	0.0534	0.0662	0.0475	0.0610	0.0434
Average(temporal)	0.0788	0.0547	0.0737	0.0523	0.0728	0.0512	0.0725	0.0514	0.0615	0.0455	0.0579	0.0425
DME	0.0691	0.0525	0.0619	0.0444	0.0643	0.0435	0.0694	0.0634	0.0624	0.0459	0.0614	0.0415
GAIN	0.0948	0.0597	0.0616	0.0509	0.0580	0.0464	0.0617	0.0491	0.0507	0.0415	0.0482	0.0390
CPH_{1-}	0.0882	0.0726	0.0513	0.0393	0.0417	0.0322	0.0624	0.0498	0.0511	0.0397	0.0365	0.0288
CPH_{2-}	0.0856	0.0678	0.0718	0.0594	0.0517	0.0371	0.0608	0.0448	0.0408	0.0324	0.0369	0.0285
CPH	**0.0573**	**0.0427**	**0.0455**	**0.0352**	**0.0392**	**0.0295**	**0.0526**	**0.0411**	**0.0400**	**0.0316**	**0.0360**	**0.0281**

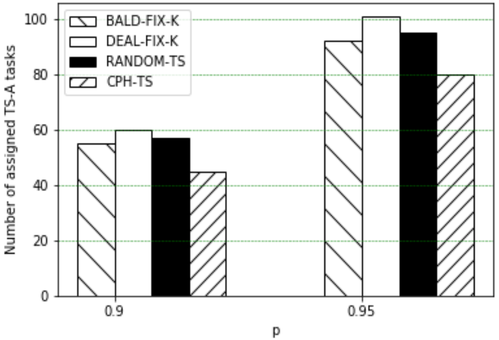
Reduction of Health Profiling Cost

Coventry University

- Assigns tasks to an average of **21.67% of regions,** while ensuring that the overall profiling accuracy meets healthcare requirement







Population health impact

Cost-effective health monitoring

- **less cost** (given a spatial coverage constraint)
- **higher** spatial coverage (given a financial constraint)

Augment existing data and address health inequality

- Improve data completeness and quality for secondary data
- know the health profiles of unknown (ignored) areas
- **Comprehensive insights and less bias** for policy making
- **Alleviate health inequality** for the overall population

Thank You !

Q&A