Heterogeneous Data Game: Characterizing the Model Competition Across Multiple Data Sources

Renzhe Xu, Kang Wang, Bo Li

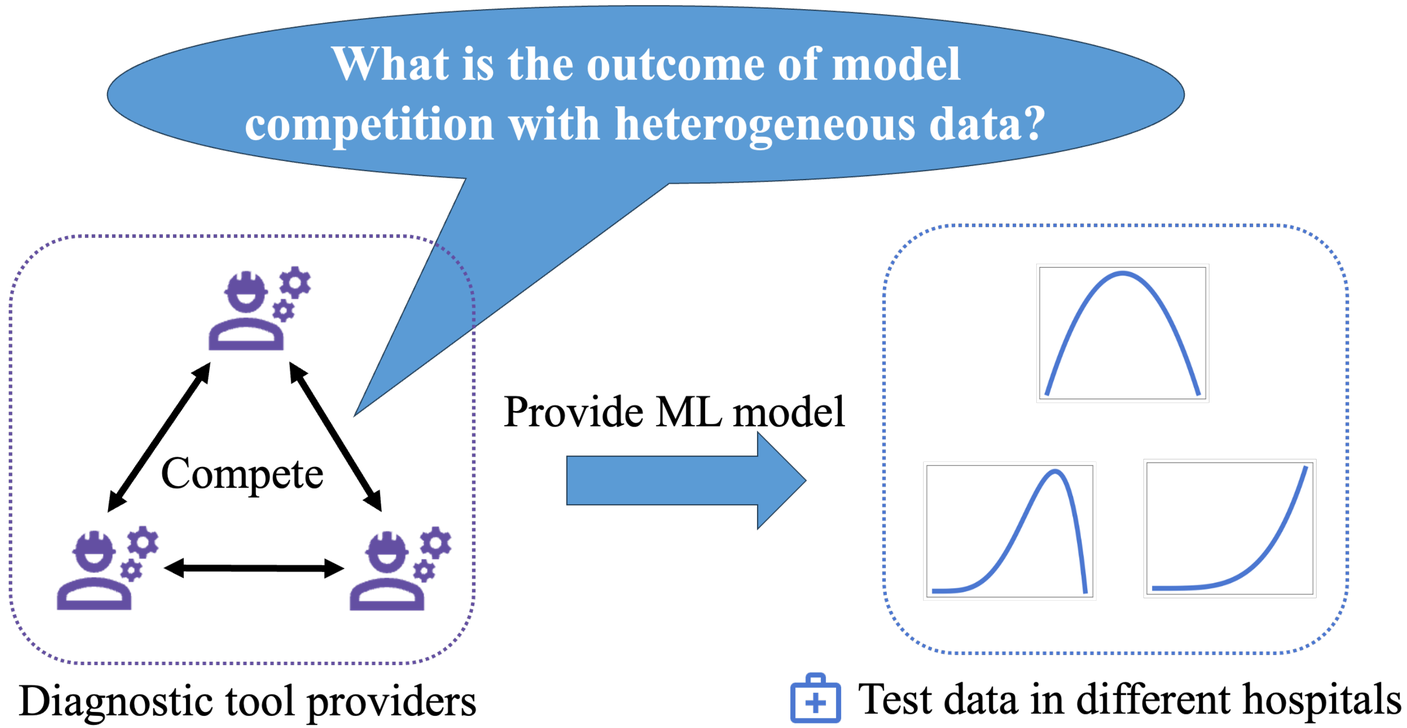
Shanghai University of Finance and Economics, Tianjin University, Tsinghua University

Background

- **Model competition with heterogeneous data**
  - **Data heterogeneity**: Machine learning (ML) models face test data from diverse data sources with varying distributions
  - **ML model competition**: Multiple ML model providers compete for market share across these data sources

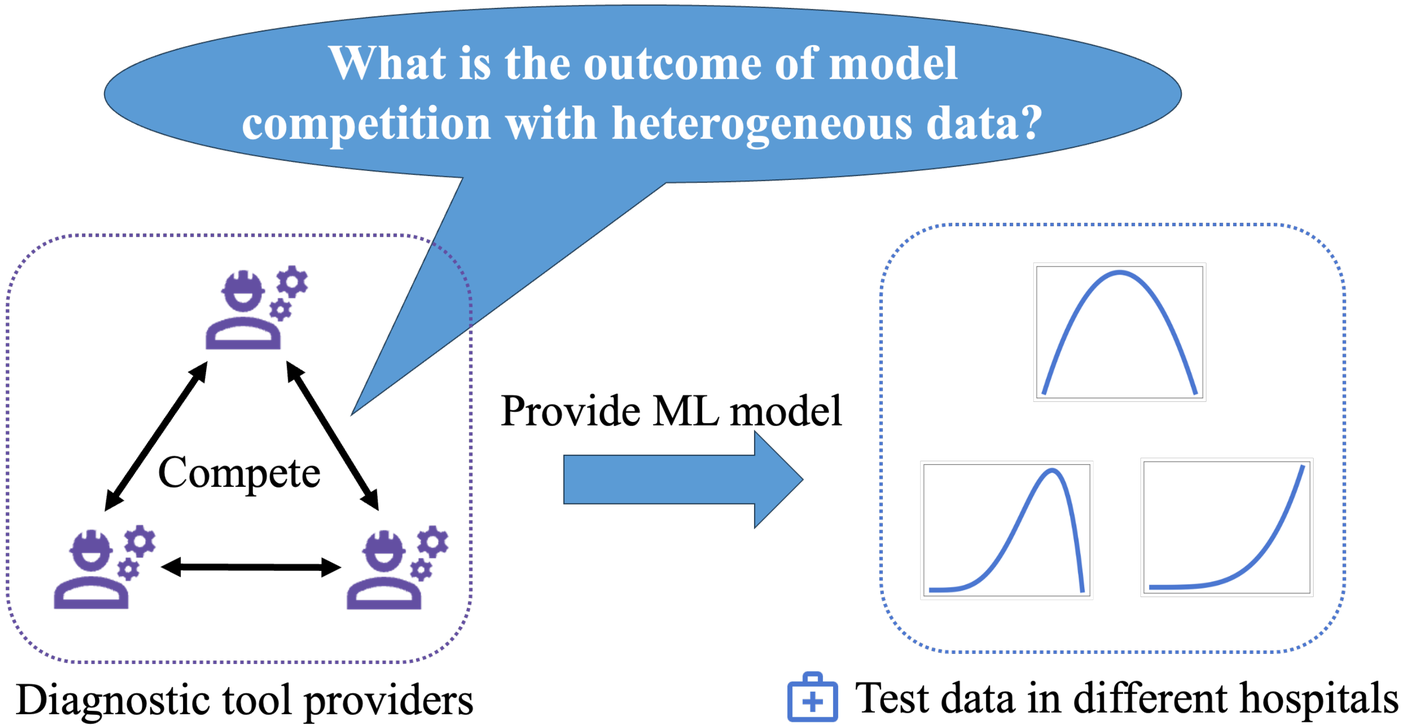
Background

- **Example in health care**
  - Different hospitals have different patient populations
  - Multiple diagnostic tool providers compete across hospitals



Background

- **Our goal**
  - What equilibrium arises in such competition?
  - What factors shape different kinds of equilibrium?



Problem setup

- **Basic parameters**
  - N model providers and K data sources
  - Each data source $k \in [K]$ has an importance weight w_k and a data distribution $P_k(x, y)$

Problem setup

- **Basic parameters**
  - N model providers and K data sources
  - Each data source $k \in [K]$ has an importance weight w_k and a data distribution $P_k(x, y)$

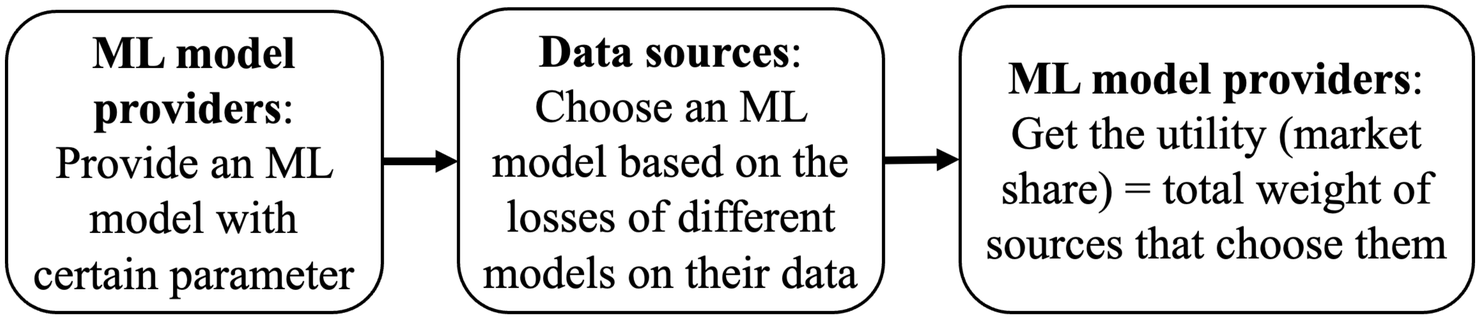
Assumption:
$P_k(y|x)$ is a linear model

Problem setup

- **Basic parameters**
  - N model providers and K data sources
  - Each data source $k \in [K]$ has an importance weight w_k and a data distribution $P_k(x, y)$

Assumption:
$P_k(y|x)$ is a linear model

- **Heterogeneous data game**



Our results

- **Three patterns of pure Nash equilibrium (PNE)**
  - **Non-existence of PNE**
  - **Homogeneous PNE**
    - All providers select the same ML model
  - **Heterogeneous PNE**
    - Providers offer different ML models and specialize in different data sources

Our results

	***Proximity* choice model** (Data sources choose the model with the lowest loss)	***Probability* choice model** (Data sources choose models via a loss-based logit function)
Monopoly (*N = 1*)	Model provider minimize the w_k-weighted loss over all data sources	
Duopoly (*N = 2*)	• Equivalent condition for PNE existence • PNE must be **heterogeneous**, if it exists	• Equivalent condition for PNE existence • PNE must be **homogeneous**, if it exists
N > 2	• Sufficient condition for PNE existence • PNE must be **heterogeneous**, if it exists	• Equivalent condition for **homogeneous** PNE existence • Sufficient condition for **heterogeneous** PNE existence • Example when both types of PNE exist simultaneously

Conclusions

- **Our contribution**
  - Model the competition among model providers in a heterogeneous data environment
  - Study how market factors affect the resulting equilibrium
    - E.g., the number of model providers, data source choice model
- **Potential policy implications**
  - Market designers could adjust market factors to achieve a more desirable equilibrium

Thanks for listening!

Renzhe Xu, Shanghai University of Finance and Economics
Paper: https://arxiv.org/abs/2505.07688
Email: xurenzhe@sufe.edu.cn