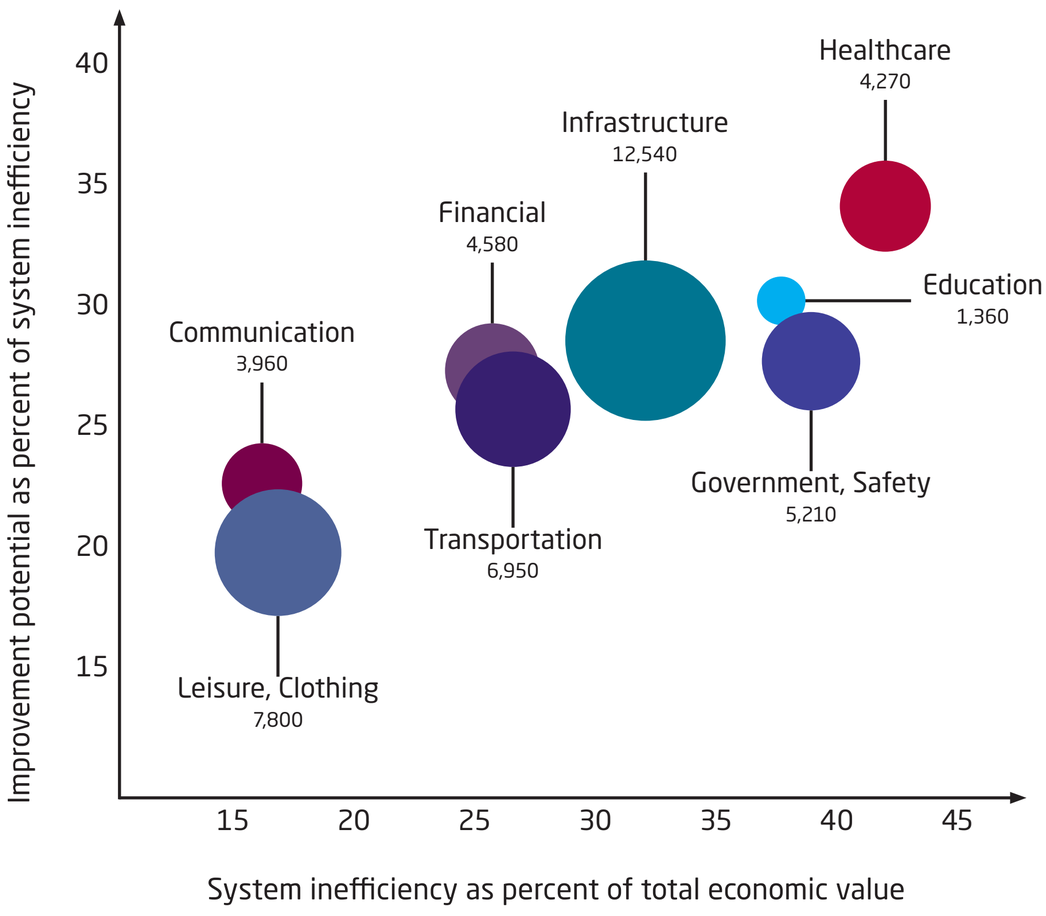
Data Engineering in Healthcare

A Market Analysis and Future Trends

Abstract

Healthcare is one of the most essential parts of modern life and we always thrive to improve it while reducing costs. Compared to other sectors, health is still relatively inefficient [1] and data engineering is one possibility to transform it in a beneficial and sustainable way. For data scientists, it is vital to incorporate domain specific knowledge, while those responsible in the healthcare sector get easily confused by buzzwords. Some of them will be explained in this poster, accompanied by a market analysis of the healthcare sector.



Market Analysis

The graph on the left (adapted from the IBM healthcare report [1], [2]) displays the relation of different sectors' inefficiency (as percentage of the total economic value) compared to the improvement potential (as percentage of the inefficiency). Furthermore, the numbers next to the sectors as well as the circle sizes indicate the absolute value of the system in billion US dollars. Healthcare has obviously one of the most inefficient systems, but the potential for improvement is also very high. This opens up a lot of entirely new opportunities. It can be concluded that:

> *The inefficiencies were attributed to several factors, including the ineffective* ***gathering, sharing, and use of information****.* [3]

Future Trends in Healthcare

Dark Data [4], [5]

Problem: *gathering of information*

- dark data = unsorted and unstructured (or even non-digital) data that has not been collected, processed, or analyzed

Solution:

- at least collect all new data
- simplify and centralize data management
- make all data accessible via centralized content repository

Data Centricity [3], [6]

Problem: *sharing of information*

- data centricity = data is the primary and permanent asset, while applications can change over time

Solution:

- technology like HL7: set of international standards about transfer of clinical data
- allow searching in the centralized content repository

Transparency [7]

Problem: (responsible) *use of information*

- false negatives can endanger patients' health, false positives lead to pricy and unnecessary treatments

Solution:

- high interpretability: no black box, as doctors are responsible for their actions
- only actionable output: e.g., age as a risk factor cannot be treated

Antje Schubotz
Master Student

Data Engineering Lecture Series
IT-Systems Engineering
University of Potsdam

References

[1] Matthias Weidlich, „Queue Mining - Analysis of Clinical Pathways based on Sensed Data", Data Engineering Lecture Series, 2017.
[2] IBM Institute for Business Value, „The world's 4 trillion dollar challenge. Using a system-of-systems approach to build a smarter planet", 2010.
[3] Manuela Müller-Gerndt, „Daten – Diagnosen – Dialog. Neue e-Health Geschäftsmodelle", 2013.
[4] Martin Grund, „Scanning Exabytes of Data using Amazon Redshift Spectrum", Data Engineering Lecture Series, 2017.
[5] Jay Savaiano, „Bring healthcare's dark data to light", 2013.
[6] Dave McComb, „The Data Centric Revolution: Data Centric vs. Data Driven", 2016.
[7] Thomas Blanchard, „Data Science and Predictive Analytics in Healthcare", 2017.

HPI Hasso Plattner Institut
IT Systems Engineering | Universität Potsdam