BIG DATA ANALYTICS IN E-HEALTH
using Apache Flink and XtreemFS

Tim Conrad, Lydia Ickler, Alexander Reinefeld,
Florian Schintke, Robert Schmidtke, Christof Schütte

Zuse Institute Berlin
Current Status & Motivation

• Managing your health in the current healthcare system is not easy. With changing health plans, multiple & disconnected practitioners, and wanting to focus on prevention, we need to take responsibility for our own health.

• Continuously collecting data allows for better health monitoring compared to that what is currently done (i.e. more or less sporadic visits to the doctor).

• We demonstrate this using blood-based analysis. More than 70% of the data needed for accurate diagnosis and health management is in your blood. It contains information about your health status, optimal disease treatment and prevention.

• One drop of blood contains 100s of GBs of data: Big Data Analytics.
Main Hypothesis

Blood contains genetic, proteomic, metabolitic and much more other information.

Diseases change signals in the blood. This means, blood samples from healthy and diseased blood samples contain differences, e.g. inflammation proteins.
Analysis Pipeline

Take the sample & process it, store it, analyze it.
Processing the Samples

A mass spectrometer produces spectra of the blood samples.

Each spectrum is about 3GB in size.
Storing the Raw Data

The raw data (spectra) can be stored transparently in a distributed fashion using XtreemFS, because to the laboratory machines it feels like a regular file system.

XtreemFS is 4x faster than using Hadoop's HDFS.
Analyzing the Data

The raw data (spectra) is analyzed in parallel using Apache Flink. Main steps include denoising, signal detection, network analysis and pattern detection (machine learning).
Key Takeaways

Modern big data-based diagnostics can help you
• keep track of your overall health in a more meaningful way and,
  more importantly,
• identify potential health threats that standard tests may not find.