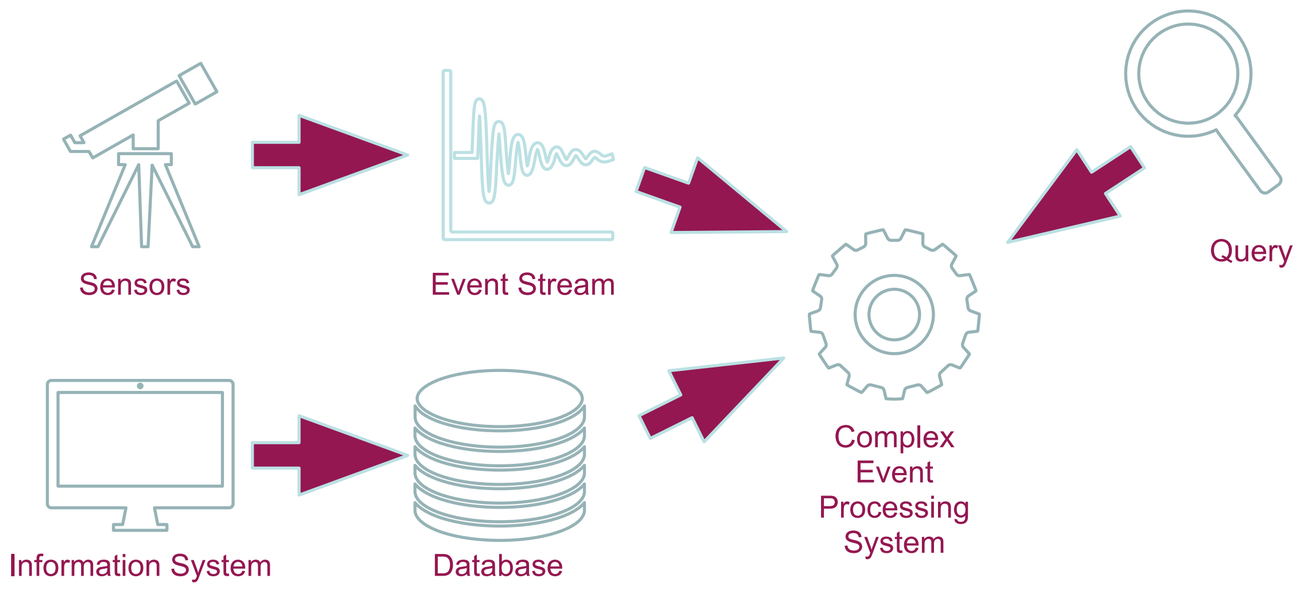
Intelligent Prefetching: Even More Efficient Complex Event Processing

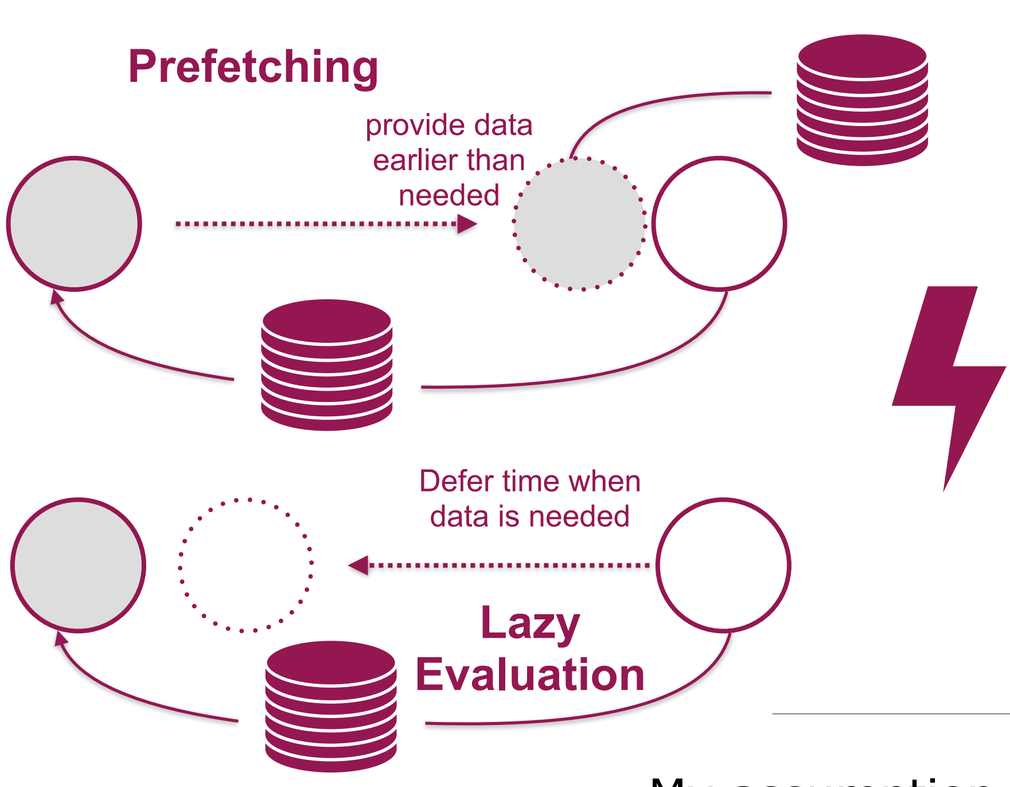
Leverage data-driven learning to optimize
complex event processing systems

Complex Event Processing comprises methods, tools and technologies to detect, analyse and process interdependent events. CEP is for example used in financial monitoring for fraud detection or prevention. The event data provided by sensors is received as event stream, for example consisting of financial transactions like bank transfers or withdrawals. Within CEP systems, the event stream is combined with data retrieved from a remote database (as depicted on the right) to answer queries.

In use cases like fraud detection or prevention **efficiency is key** as the value of analyses decreases over time. The goal of this research proposition is to enhance current CEP systems in terms of efficiency.

The proposition is to contribute to the group of Matthias Weidlich at Humboldt University in Berlin, who already developed a series of optimizations concerned with the different parts of CEP.



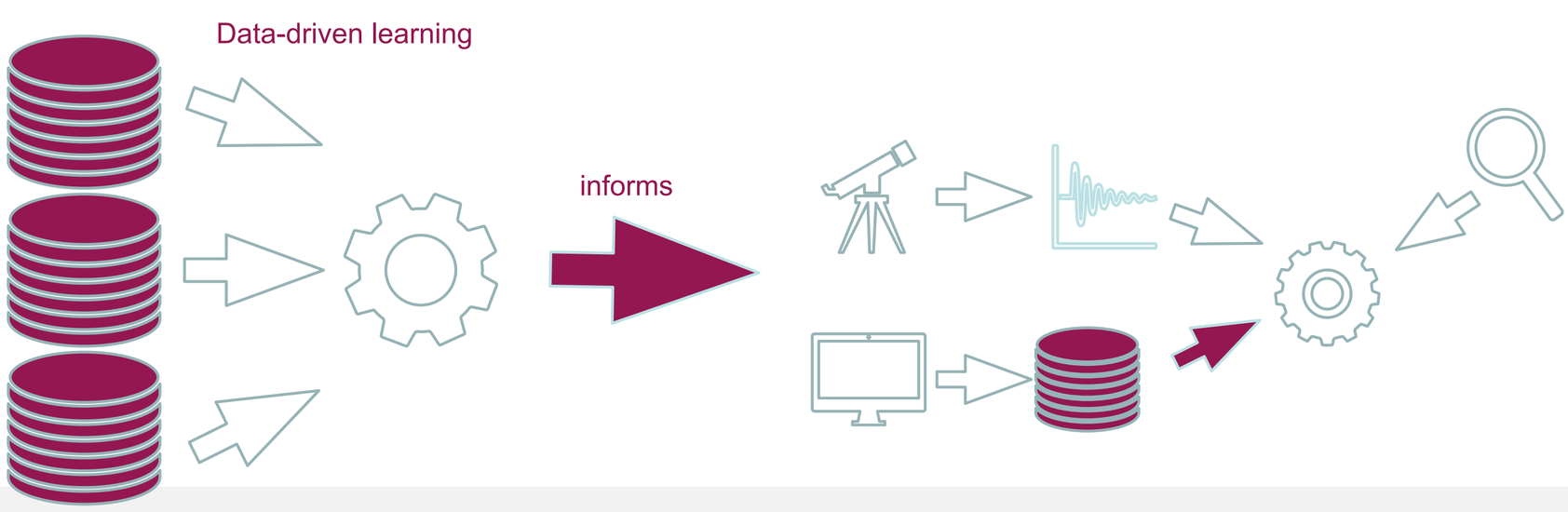


One proposed optimization approach addresses efficient remote data integration by using a hybrid approach of lazy evaluation and prefetching. The combination aims at balancing between lowering the transmission latency and making more inaccurate predictions, resulting in additional matches. The decision about which data to prefetch is based on cache miss statistics maintained by the framework. However, as cache miss statistics are rather reactive, they might not be sufficient in any case. Further, by only relying on cache miss statistics, we don't **leverage knowledge** about the data retrieved from the remote database, that could be available **prior to the query** execution.

My assumption is that we can enhance prefetching by going beyond relying on cache miss statistics and instead leveraging multivariate data distributions through data-driven learning as presented by Binning and his group at TU Darmstadt to inform the framework. These are currently used to improve join orders by estimating cardinalities.

The multivariate data distributions could also be used to **optimize prefetching**.

Making better predictions and therefore informed decisions about which data should be prefetched, should **minimize transmission latency** and simultaneously not lead to additional matches.



Caterina Mandel, M.Sc. IT-Systems Engineering

Hasso Plattner Institute, Potsdam, Germany

E-Mail: caterina.mandel@student.hpi.de

Hasso Plattner Institut