

# **AEGLE'S Brief nº1**

## First system design including system requirements

By Certh, Croydon, Exus, Gnubila, ICCS, Kingston, Maxeler, Pagni, Uppsala University, Universita San Raffaele.

AEGLE targets to build an innovative ICT solution addressing the whole data value chain for health based on: cloud computing enabling dynamic resource allocation, HPC infrastructures for computational acceleration and advanced visualization techniques and contribute in the area of big data analytics for Health Bio-data.

Following the definition by the IMIA working group on "Data Mining and Big Data Analytics", "Big Data are data whose scale, diversity, and complexity require new architecture, techniques, algorithms, and analytics to manage it and extract value and hidden knowledge from it".

To put this work in perspective, big data in health refers to electronic health-related data sets that cannot be managed with traditional software and/or hardware and common methods. Big data is bringing challenges to traditional data processing, as regards the size of data (volume), the required processing speed (velocity), the heterogeneity (variety) and the accuracy (veracity).

Indeed health data volume is expected to grow dramatically and even now in its totality is overwhelming, including different types and variable analytics scope:

 a) EHR longitudinal data useful for predictive analysis and deep phenotyping, along with multimedia clinical data medical imaging, laboratory, pharmacy, insurance;

b) Personal quantifiable and social data in telehealth, social media

and quantified-self domain;

c) Omics data for precision medicine and individualized treatments

d) medical knowledge and literature.

Although ICT makes advancements to address volume and velocity of big data, and health big data applications have been recently proposed, the healthcare industry has been hesitant in embracing big data. AEGLE aims to generate value from healthcare data with the vision to improve translational medicine and to facilitate personalized and integrated care services overall improving healthcare at all levels, to promote data-driven research across Europe and to serve as an enabler technology platform. Privacy, openness and sharing concerns are raised health-related data. Their expected heterogeneity and poor quality has to be overcome for veracity, while paradigm shifts towards data-driven analytics have to be accepted. AEGLE aims to address these bottlenecks and contribute towards filling these gaps in a range of medical problems, concerning electronic health record data, biological data and streaming biosignal data.



#### **Overview of the three AEGLE use cases**



The most crucial challenge for the success of big data in health is to make value out of these data. Health research has been built on small and clean data, with carefully designed cleaned trials and extrapolation of their findings. A shift from hypothesis-driven to data-driven research is foreseen, based on machine learning techniques that mine patterns, clusters and associations for big (e.g. population representative) volumes of unclean data. To increase medical credibility, the produced knowledge and hypotheses can then be confirmed in smaller and cleaner datasets. The three medical use cases of AEGLE have been carefully chosen to cover biomedical research and questions that can set the basis for biosignal and bioinformatics analytics, multiparametric pattern mining, and integrative predictive modelling. Presentation of information and visualisation techniques for a multitude of medical data types, and their interconnections, will be another complexity level.

#### The five dimensions (5's) of Big Data in AEGLE cases

The challenges include capturing, storing, searching, sharing & analyzing.

Case	Volume	Velocity	Variety	Veracity	Value
CLL	High-Throughput Sequencing Data			Error-Prone Data Generation	Understanding disease pathophysiology / risk stratification / novel therapeutic targets
ICU		Streaming Data		Artifacts	Predict / Alert / Intervene
Diabetes			Diverse EHR Structures and coding schemes	Missing Values	Estimate risk / Optimize Treatment

All the three medical cases include clinical data, and thus management and analysis of structured clinical data may follow a common logic, at least as regards storage, exploratory analysis and visualisation. Both Diabetes and CLL case address chronic diseases ICU case, on the other hand, addresses acute care questions. In this case, clinical data will be combined with features derived from biosignals (e.g. cardiac signals) and the medical equipment (e.g. ventilation use), and thus machine learning techniques for biosignal processing is expected to be more intense. It has to be noted that quality issues, error correction, adjustments for missing values, or data bias, will be dealt all cases, to increase the credibility of the AEGLE analytics, and result in actionable results.

AEGLE starts from real life conditions in the three medical cases,

which include clinical and biomedical research. It is thus important to consider not only the data that are produced at each organization, but also the procedures and analysis pipelines that are in place. Therefore AEGLE defines the local and the cloud domain, as well as a loose coupling between them. The former refers to the organizational private space, where routine use of data takes place, as well as analysis without privacy concerns.

The latter refers to the cloud space, where data from multiple organisations can be shared, combined and analysed, provided the necessary access and privacy-related preprocessing. Biodata repurposing and analytics at cloud level can serve in new knowledge discovery, that will then be useful at each medical domain.



Finally, in AEGLE the path from data to knowledge, interpretation, actionable data, necessarily involves open data standards for sharing and interoperability, methods for semantic and temporal similarity, as well as standardized integration of data, e.g. clinical

and genomic. Applying, interlinking and extending current medical standards for the integrated and quality based use or even repurposing of big biodata will be a key issue in AEGLE, addressing both variety and veracity.

### High-level illustration AEGLE cloud platform.

Along with the considered technological artifacts.



The Aegle Project Aegle's mission is to realize an European business ecosystem to healthcare stakeholders, industry and researchers for creating out-of-box knowledge in order to provide cloud and HPC data services and support new products that will improve health.