de.NBI
GERMAN NETWORK FOR BIOINFORMATICS INFRASTRUCTURE

Annual Meeting of the
de.NBI Industrial Forum
Thursday, 23 November 2023

Book of Abstracts – Scientific Presentations

November 23rd 2023, 01:00-05:00 pm (CET)

Virtual Meeting using Zoom

Log-In data will be sent to you after the registration

Session A - Cloud Computing for the analysis of Life Science data

Introduction to the de.NBI Cloud: Organization, Governance, and Use Cases

Alexander Sczyrba
Director IBG-5, Institute of Bio-and Geoscience, Computational Metagenomics
de.NBI Cloud Coordinator, Forschungszentrum Jülich GmbH

Abstract:

In recent years, the rapid technical improvements in modern life science research led to the generation of huge amounts of experimental data. To meet these rising demands, the German Network for Bioinformatics Infrastructure (de.NBI) was established to provide high quality bioinformatics services, training, computing capacities (de.NBI Cloud) as well as connections to the European Life Science Infrastructure ELIXIR, with the goal to assist researchers in exploring and exploiting data more effectively.

The establishment of the de.NBI Cloud has proven to be a flagship for the de.NBI network. It consists of eight federated cloud locations that implement a common governance and use the project application and management workflow provided by the de.NBI Cloud portal. This governance facilitates secure operations of each cloud location through the centralized organization of ISO27001 Information Security Management System (ISMS) training and certification courses for the cloud staff, leading to the progressive certification of our cloud sites. The de.NBI Cloud portfolio includes several project types designed to suit different use cases and users with varying levels of knowledge in cloud computing. Two project types, OpenStack and Kubernetes, offer maximum flexibility in terms of the configuration of cloud-specific components and allow the installation of any large-scale analysis, stream processing or orchestration framework available in the cloud ecosystem. Both project types are ideal for science gateway developers to offer bioinformatics services to the national and international life sciences communities. Confidential processing of sensitive data, e.g. pseudonymized patient-related data, is also possible at specific de.NBI Cloud locations, where data security is enforced through separated, secure processing environments.

Our in-house developed project type SimpleVM enables our users to use cloud resources with little to no background knowledge in cloud computing. SimpleVM is an abstraction layer on top of OpenStack to manage single virtual machines (VMs) or clusters thereof. It was designed to support the combination of resources from independent OpenStack installations, thus operating as a multi-cloud platform which is accessible from a single web-based control panel. In general, SimpleVM primarily eases the creation and management of individual pre-configured virtual machines and provides web-based access to popular research and development environments such as Rstudio, Guacamole Remote Desktop, Theia IDE, JupyterLab and Visual Studio Code.

NFDI4Microbiota: Advancing (Big) Data Solutions in Microbiology

Konrad Förstner
Head of Data Science and Services
ZB MED - Information Center for Life Sciences, Cologne

Abstract:

NFDI4Microbiota (The National Research Data Infrastructure for Microbiota) represents a pivotal initiative within the National Research Data Infrastructure (NFDI) framework, dedicated to advancing (big) data applications in life sciences, specifically in the realm of microbiota research. The initiative aligns with national and international data management standards, ensuring compatibility and harmonization with broader research initiatives and aims to make research data open and FAIR (findable, accessible, interoperable, reusable). It is/will be offering data storage, analysis services, training, consultancy and much more for the microbiology research community. Ultimately, NFDI4Microbiota aspires to accelerate discoveries in microbiota research, contributing to our understanding of the intricate relationships between microorganisms and their environments through seamless data management and analysis.

HEALTH-X dataLOFT: Vision of a European Health Data Platform

Harald Wagener
Group Leader Cloud
Berlin Institute of Health, Charité, Berlin

Abstract

Since two years, the HEALTH-X dataLoft project, supervised by the Center for Digital Health at BIH@Charité, develops a Gaia-X based platform realizing an open and federated data transfer. This data transfer has been legitimized by the public. The presentation given by Harald Wagener provides an overview of the project’s current status and progress so far as well as an outlook in regards to HEALTH-X dataLOFT in the context of the European Health Data Space (EHDS) and Co.

Session B - Analysis of Life Science data in the Industry

Navigating Zillions of Molecular Possibilities in Early Therapeutic Drug Design — The Ultralarge Chemical Space Revolution

Marcus Gastreich
SVP Application Science
BiosolveIT GmbH, Sankt Augustin

Abstract

In the realm of early drug design, we are currently witnessing two profound transformations: the integration of AI and the exponential growth of Big Data. Starting from a notable collaboration with Enamine, a prominent Ukrainian Contract Research Organization (CRO), BioSolveIT has amassed market-dominating expertise in navigating vast quantities of molecules. This partnership enables lightning-fast exploration of chemical spaces, swiftly identifying molecules with desirable properties that, by definition of the process, can be made and purchased.

At the core of this process are speedy and new algorithms, empowering researchers to explore trillions of molecules within seconds, on standard hardware. This leads to a drastic reduction in time and costs when searching for promising drug candidates. Furthermore, the availability of purchasable molecules, facilitated by partnerships like that between BioSolveIT and Enamine — and over time also other CROs such as WuXi, OTAVA, or marketplaces such as eMolecules — accelerates the entire development process. Merck, for example, have published savings of 90%.

Today, chemical space navigation across the synthetically accessible space is an indispensable tool for the pharmaceutical industry. The success from off-the-shelf molecules compels pharmaceutical companies to develop their own chemical spaces and mine from incomparably larger pools of intellectual property. An interesting step even further consists in coupling and integrating these technologies with robotics to stay competitive.

The talk shall give a broad overview of these developments, sketch up the basic ideas of the search technologies, and showcase its capabilities with a few success stories from the industry

Large cohort analysis for 4D-Metabolomics, Lipidomics and Proteomics

Nikolas Kessler
Head of Software R&D Metabolomics/Lipidomics
Bruker Daltonics GmbH & Co. KG, Bremen

Abstract

Mass spectrometry has emerged as a powerful analytical tool in the fields of proteomics, metabolomics, and lipidomics, enabling the comprehensive analysis of biomolecules within complex biological systems. Bruker's timsTOF series of instruments, coupling liquid chromatography (LC) with trapped ion mobility spectrometry (TIMS) and mass spectrometry (MS) with parallel accumulation serial fragmentation (PASEF®) provides extremely high MS/MS speed and sensitivity. This enables 4D-Omics for deep and complete characterization of samples in a high-throughput fashion.

The global profiling of small molecules in complex and large cohorts of biological samples using these techniques produces complex multidimensional data for hundreds to thousands of metabolites or lipids. Here we present the RealTimeQC module of the TASQ® software for in-depth analysis of data quality during acquisition, and Bruker's complete QC workflow which includes the MetaboScape® software for post-acquisition review and correction of profiling data. MetaboScape's REST API furthermore allows full integration of its data processing and molecule annotation capabilities into streamlined bioinformatics pipelines.

Finally, with Bruker ProteoScape™ we present a GPU-powered platform delivering parallel computing capabilities across thousands of CUDA cores and real-time database search results for bottom-up proteomics, making sure meaningful data is generated from precious samples

Building flexible and scalable infrastructure for data-driven synthetic biology using Cloud Computing

Jaqueline Hess
Cambrium GmbH, Berlin

Abstract

Data-driven synthetic biology requires significant computational and storage resources as well as the appropriate tooling to ensure FAIR principles for data generation across wet and dry labs and version controlled pipelines. At the same time, it is imperative to retain the ability to rapidly prototype and deploy new applications within an existing ecosystem. Cambrium is a data-driven biomaterials company developing protein-based materials to replace animal and petrochemical-derived products from our supply chains. Our platform harnesses protein design, generative AI and machine learning to design high-performing proteins with specific functions, improve their expression and scale their production. Here we show how Cambrium uses a combination of open-source and proprietary cloud-based tooling to power our data backend and move from raw data to actionable results quickly and with minimal hands-on time. We will use our NGS processing pipeline as an example to demonstrate data integration across different data layers and how we can now move from DNA extract to characterised strains within 24 hours.

The Impact of Artificial Intelligence in Clinical Research

Steven Lazer
Global Healthcare & Life Sciences, CTO
Dell Technologies

Abstract

The appetite of Artificial Intelligence to consume data is insatiable. Creating and accessing large data sets for clinical research can be a daunting task. The task of data curation alone represents a difficult challenge. Data sets need to be clean and deidentified to create those opportunities for data sharing while eliminating challenges with data sovereignty. Come understand methodologies to create and manage data sets to support clinical research with the opportunity of combining global data sets to create a SDOH sets of democratized data. Understand data access methodologies, synthetic data opportunities, and bringing research to the data to create access to data sets you do not own. Once the datasets become available the question comes as to what to do with them. New advances in artificial intelligence and data processing capabilities with alternative processors establish a pathway for things like a clinical digital twin and advanced research.

Big Data Handling in the Life Science: MAKING DATA ACTIONABLE.

Andreas Kremer
Co-Founder and Managing Director Information Technology for Translational Medicine (ITTM) S.A., Luxembourg

Abstract

Providing access to data for scientific research purposes is a common scenario in many projects. This has, in recent times, gained increasing importance and complexity. Additionally, a data sharing culture shift is needed within organizations, where the added value of data sharing is understood, and inner resistance is overcome. There is a cost of not cooperating – not sharing: delaying break-through discoveries, missing out on cost savings by duplicating efforts, or simply the risk of there being positive outcome/ result at all. To quote: “Co-opetition (cooperation with competitors – academic or industry) requires mental flexibility, but groups that develop it can gain an important edge.” Doing this in a safe sandbox environment, such as institutional public private partnerships, allows to capitalize on sharing and make sure that interests of all parties are taken care off.

BIG DATA handling in a Plant Breeding Company

Andreas Menze
RD DataScience & Analytics
KWS Saat SE, Einbeck

Abstract

Plant breeding is a complex and time-consuming process. To develop new plant varieties, breeders need to collect and analyze large amounts of data. Efficient **Big Data handling** help breeders to speed up this process.

With respective **data analytics on BIG DATA** breeder can enhance knowledge gaining new insights into plant physiology and development, leading to new varieties with improved traits like yield, resistance to diseases or drought.

Big Data handling in a plant breeding company comes with a number of **challenges** according typically to the generally "3 V's" characteristics: **Volume** referring to the sheer amount of data like the increasing number of sensor and molecular data. **Variety**: in plant breeding, data is generated from a variety of sources, such as lab tests, greenhouse- and field trials providing all kind of complex data for phenotypic, molecular, and environmental information. **Velocity**: Data is often captured in real time and/or it requires the ability to deal with data processing in real-time to breeder. Additionally **Veracity** is highly important dealing with reliability and quality of data.

To overcome these challenges different measures and aspects have to be considered. This concerns a well thought through IT architecture dealing with structured and unstructured data, centralized and decentralized solutions being well connected to data analytics. This is supported by a Data Governance framework.

Empowering insights in KNIME: from Big Data to Large Language Models

Martyna Pawletta,
KNIME GmbH, Berlin

Abstract

With the tremendous growth of data in many disciplines, big data has already been a topic in the data science world for a while. With new tools coming to the forefront every year, including different cloud environments and new technologies, analyzing and extracting knowledge from big data sources continues to become easier.

Especially in healthcare as well as life sciences. With the high diversity of data types and use cases coming from research, the development of new technologies is even more important. Years of attention within the area of Text Processing and NLP make textual data especially relevant. While big data still remains a “big” topic, another technology has become the new “big” - or rather “large” for Large Language Models (LLMs).

In this presentation I would like to introduce different capabilities of the KNIME Analytics Platform, an open source tool for low-code data science. I will focus on features around big data, data engineering and its recently implemented AI capabilities. This opens new possibilities, especially for the community working in research as it provides easy access to data science, with a reduced barrier to start experimenting with new models available for Life Sciences.

Empowering Life Science Innovation Through Integrated Big Data Management: From Instruments to the Cloud with FAIR Data

Maximilian Wiens,
Wiens Synefex GmbH, Leopoldshöhe

Abstract

The large field of life sciences is increasingly dependent on the sophisticated management of big data, necessitating advancements in integration techniques from initial data acquisition to advanced cloud-based analytics. This talk provides a systematic examination of the lifecycle of big data within this domain, adhering to the foundational FAIR data principles. It addresses the methodological basis of device integration for comprehensive data capture and the subsequent necessity of metadata enrichment workflows. The talk outlines methods for secure and scalable data storage, alongside strategies for distributing data across diverse systems. It examines the use of data lakes and data marts for effective data governance, emphasizing the critical importance of standardized data definitions and structures for interoperability. Highlighting the necessity of advanced analytical tools for the visualization and understanding of complex data sets, it argues for strategic data management as a driver of innovation in the life sciences. The synergy between technical expertise and scientific investigation is showcased, advocating for an integrated approach to big data management as essential to contemporary life science research.