*Die hier skizzierte Idee zu einem Kompetenzzentrum für Digitalisierung und Data Science im experimentellen Setting ist eine Idee, die im kleinen Kreis der hier genannten Autor*innen entstanden ist. Es soll Grundlage und Anstoß für die weitere Diskussionen sein an dem alle Lesenden eingeladen werden sich zu beteiligen.*

Pilotierung eines Kompetenzzentrum für Digitalisierung und Data Science im experimentellen Setting

Autor*innen/Initiator*innen:

Prof. Dr. Peter Neubauer, Institut für Biotechnologie, Fachgebiet Bioverfahrenstechnik, Technische Universität Berlin

Dr. Marie-Therese Schermeyer, Institut für Biotechnologie, Fachgebiet Bioverfahrenstechnik, Technische Universität Berlin

Nicolas Cruz Bournazou, Institut für Biotechnologie, Fachgebiet Bioverfahrenstechnik, Technische Universität Berlin

Veronique Riedel, Präsidialamtsleiterin, Technische Universität Berlin

Prof. Dr. Geraldine Rauch, Präsidentin, Technische Universität Berlin

Bestätigte und anzufragende Partner*innen (Liste erweiterbar, Vorschläge willkommen):

HU Berlin:

Prof. Dr. Ulf Leser, Wissensmanagement in der Bioinformatik, Mathematisch-Naturwissenschaftliche Fakultät, Humboldt Universität zu Berlin (bestätigt)

Prof. Dr. Matthias Weidlich, Datenbanken und Informationssysteme, Mathematisch-Naturwissenschaftliche Fakultät, Humboldt Universität zu Berlin (bestätigt)

FU Berlin:

Prof. Dr.-Ing. Jochen Schiller, AG Technische Informatik, Institut für Informatik, Fachbereich Mathematik und Informatik, Freie Universität Berlin (bestätigt)

Charité:

Prof. Dr. Andreas Thiel, Regenerative Immunology and Aging, BIH Immunomics, Charité Universitätsmedizin Berlin (bestätigt)

Prof. Dr. Birgit Sawitzki, Institut für Medizinische Immunologie, CVK, Südstraße 2, Charité Universitätsmedizin Berlin (bestätigt)

Prof. Dr. med. Dr. rer. nat. Felix Balzer, Institut für Medizinische Informatik, Charité – Universitätsmedizin Berlin (bestätigt)

Prof. Dr. Frank Konietschke, Institut für Biometrie und Klinische Epidemiologie, Charité – Universitätsmedizin Berlin (anzufragen)

Prof. Dr. Tobias, Institut für Public Health, Charité – Universitätsmedizin Berlin (anzufragen)

TU Berlin:

Prof. Dr.-Ing. Lydia Kaiser, Institut für Werkzeugmaschinen und Fabrikbetrieb, Fachgebiet Digitales Engineering 4.0, Technische Universität Berlin (bestätigt)

1. Ausgangssituation und Hintergrund

Im Hinblick auf wachsende Datenmengen in Laboren und experimentellen Setting und dem Anspruch der Wissenschaft hin zu fachübergreifendem Arbeiten, sehen wir den dringenden Bedarf Infrastruktur, Methoden und Wissen im Bereich Data Science verstärkt in die naturwissenschaftlich - empirische Wissenschaft und akademische Ausbildung zu integrieren.

Aktueller Stand:

- In den meisten Laboren bzw. Projekten erfolgt die Planung, Durchführung und Auswertung der Experimente traditionell. Die Daten werden personenabhängig entsprechend lokaler Vorgehensweise oft in persönlichen oder institutionellen Serverbereichen abgelegt. Gerätespezifische Datenbanken enthalten oft nur die Daten der spezifischen analytischen Methode des jeweiligen Gerätes und nicht die Metadaten der durchgeführten Experimente, die für das Verständnis der Argumentation (Warum wurde das Experiment genauso durchgeführt? Welche Spezifika gab es?) und für eine exakte Reproduktion notwendig ist.
- Die derzeitige Vorgehensweise erlaubt es nicht, die detaillierten Informationen, die zu jedem Experiment und seinen Ergebnissen gehören, innerhalb der Einrichtung oder nach außen in einfacher, strukturierter und vielseitig anschlussfähiger Form an Forschungspartner weiterzugeben.
- Obwohl die Datenverwaltung nach den FAIR-Prinzipien von den Fördergebern gefordert wird, wird dieses nur sehr limitiert und meist in Bezug auf ausgewählte Datensätze der final publizierten Ergebnisse umgesetzt, so dass Experimente dann so auch nicht in anderen Laboren reproduziert werden können. Bei der Neu- und Weiterentwicklung einer Core Facility müssen Strategien entwickelt und implementiert werden, die folgende Elemente enthalten: (1) Analyse und Automatisierung von Laborarbeitsabläufen, (2) Entwicklung vo Datenbanken und Austauschformaten in denen sowohl die Experiment-spezifischen Daten als auch die gesamten Metadaten in einer Form abgelegt werden, die einen reibungslosen Austausch gestatten und in der auf diese mit verschiedenen Auswerteprogrammen, Simulationsmodulen etc. zugegriffen werden kann. Die Gesamtdaten eines Experiments bzw. Projekts können vollumfänglich auch Partnern zur Verfügung gestellt werden. (3) Anbindung von technischen Geräten, um Workflows transparent zu machen und logisch zu verbinden.
- Innerhalb der Berlin University Alliance gibt es herausragende Forschungsaktivitäten an allen Häusern im Bereich Data Science/Data Management/KI/Big Data. Allerdings ist die Implementierung der Kompetenzen in die Landschaft der Forschungslabore und naturwissenschaftlichen Forschungsprojekte noch eine Herausforderung.
- Es fehlen strukturierte Lehr- und Weiterbildungsangebote sowohl auf studentischem als auch auf postgraduiertem Niveau, um angehenden Forscher*innen der experimentellen Wissenschaften die grundlegende Prinzipien der Digitalisierung, der Datenwissenschaften und des datenanalytischen Denkens zu vermitteln.

- Es gibt zwar bereits vereinzelte Service Angebote im Bereich Data Science an den Häusern. Diese sind aber unterbesetzt und leiden unter mangelnder Verzahnung, was Synergien verhindert. Es gibt aber keine zentrale Anlaufstelle für Nutzer*innen, wo der Leistungsbereich existierenden Core Facilities zusammengefasst und dargestellt wird.

Existierende Initiativen:

- An den Häusern gibt es bereits verschiedene **Core-Facilities**, die Dienstleistungen im Bereich Data Science anbieten, z.B.
    - Für eine bessere Verzahnung der Forschungseinrichtungen ist der Bau eines neuen Forschungsgebäudes – „der Simulierte Mensch (Si-M)“ begonnen. Hier werden Ingenieure und Kliniker der verschiedenen Entwicklungsebenen aus Charité und der TU Berlin räumlich zusammengebracht, um Grenzen zwischen den unterschiedlichen Wissenschaftskulturen zu überwinden. Die Schaffung gemeinschaftlich genutzter Labore und Analytik muss jedoch auch von einer entsprechenden IT Infrastruktur unterfüttert werden um erfolgreich zu sein. In dem am Si-M initiierten Pilotprojekt, das bereits als Core Facility etabliert wird, werden Erkenntnisse gewonnen und Methoden der Digitalisierung und Automatisierung etabliert.
    - DAS KIWI-biolab-biolab „Internationales Zukunftslabor für Künstliche Intelligenz in der Bioprozessentwicklung“ innerhalb des Si-M als gemeinsames Pilotprojekt die datenbasierte effiziente und intelligente Vernetzung von Laboren für die Partner*innen der Berlin University Alliance zu etablieren.
    - Das Servicezentrum Forschungsdatenmanagement an der HU Berlin,
    - Das Servicezentrum Forschungsdatenmanagement an der TU Berlin,
    - An der HU beraten zu Fragen der Data Science verschiedene Verbundprojekte, wie der SFB 1404 "FONDA – Foundations of Workflows for Large-Scale Scientific Data Analysis" oder die gerade neu eingerichtete Forschungsgruppe „Integration von Deep Learning und Statistik zum Verständnis strukturierter biomedizinischer Daten“
    - Verbundübergreifend und unter Einbeziehung der sechs Helmholtz Zentren in Berlin und Umgebung bietet das domänenübergreifende Graduiertenkolleg "HELMHOLTZ EINSTEIN INTERNATIONAL BERLIN RESEARCH SCHOOL IN DATA SCIENCE (HEIBRiDS)" Unterstützung zu Fragen der Anwendung von Data Science in den Naturwissenschaften.
    - Die Core Facility CUBVI am BIH / der Charit´´e Berlin, die Forscher*innen der Charité in Fragen der Datenspeicherung, -integration, und- auswertung vor allem in der Genomik unterstützt,
    - Die Service Unit Biometrie der Charité – Universitätsmedizin Berlin, die das Beratungsdienstleitungen für Datenauswertungen kostenfrei anbietet.

- An den Häusern gibt es bereits verschiedene Lehr- und Weiterbildungsangebote:
    - An der FU Berlin gibt es einen konsekutiven Masterstudiengang „Data Science“, der aktuell die beiden Schwerpunkte „Data Science in Life Sciences“ und „Data Science Technologies“ umfasst und damit von Studierenden aus den Naturwissenschaften wie auch Mathematik und Informatik gewählt werden kann. Die Erweiterung auf weitere Fachgebiete ist in Vorbereitung.
    - An der HU Berlin ist ein Masterstudiengang „Data Science“ nach Typ 1 (gemäß GI Klassifikation) geplant, in dem Studierende der Informatik oder Mathematik

vertiefende Kenntnisse in Maschinellem Lernen und Data Engineering erfahren werden und sich in einer naturwissenschaftlichen Anwendungsdomäne vertiefen können.
- An der Charité – Universitätsmedizin Berlin gibt es einen PhD Studiengang „Health Data Science".
- An der TU Berlin wurde an der Fakultät IV die studienübergreifende Vorlesung „Data Science 1" nach dem Vorbild der „Foundations of Data Science" Vorlesung an der UC Berkeley initiiert.

Die Einrichtung einer BUA-weiten Initiative und Koordination zur Förderung von Digitalisierung und Data Science insbesondere für naturwissenschaftliche und experimentelle Fächer wird von allen Häusern gestützt. Es gibt insbesondere Chancen und Bedarfe in der Ausbildung, und dort sowohl auf dem Level von Postgraduierten als auch bei Master-Studierenden. Da die Häuser unterschiedliche Forschungsschwerpunkte mitbringen, kann eine solche Initiative vom Verbund nur profitieren.

Die vier Verbundpartnerinnen der Berlin University Alliance (BUA) haben sich den gemeinsamen Aufbau und die Stärkung nachhaltiger Data Science Strukturen zum Ziel gesetzt. Zentraler Gegenstand ist dabei die Konzeption eines Netzwerks institutionenübergreifender Data Science Services. Dies kann durch ein Kompetenzzentrum für Digitalisierung und Data Science im experimentellen Setting realisiert werden.

Das Netzwerk an kooperierenden Gruppen soll dabei nach einer Pilotphase stetig erweitert werden, und zwar (a) innerhalb der Berlin University Alliance, (b) auf andere universitäre und außeruniversitäre Partner*innen und (c) auf deutschlandweite und internationale Kooperationen mit akademischen und industriellen Partnern.

2. Aufgaben eines Kompetenzzentrums für Digitalisierung und Data Science im experimentellen Setting

Das geplante Kompetenzzentrums für Digitalisierung und Data Science im experimentellen Setting soll folgende Aufgaben haben:

- Nutzer*innenschnittstelle im Sinne eines Single Point of Entry mit Bündelung existierenden Services und zur Etablierung neuer Services im Rahmen eines fächerübergreifenden Core Facility Netzwerkes
- Experimentierplattform zur Entwicklung bzw. Erprobung neuer Methoden und Strategien zur Datenablage, zum Datenhandling, wie zur Datennutzung z.B. in intelligent gesteuerten Experimenten
- Koordination und Vernetzung mit internen und externen Partner*innen aus dem (außer-)universitären und industriellen Bereich
- Schnittstelle zum Zugang zu Hochleistungsrechner-Ressourcen an den Häusern oder am Zuse Institut Berlin
- Einrichtung und Organisation einer verbundübergreifenden Graduiertenschule „Data Science", die als synergistischer Dachverbund für existierende Graduiertenkollegs angelegt wird (wie z.-B: GRK Heibrids, IGRK 2403 Analyse des regulatorischen Genoms, GRK 2424

CompCancer, GRK 2434 „Facets of Complexity“, ExIni-GRK SALSA – Analytical Sciences, etc.) und zentrale Bereiche wie Rekrutierung, Weiterbildung, oder On-Boarding ausländischer Promovierender übernehmen wird.

- Aufbau von fachspezifischen Best-Practices Datenbanken
- Abstimmung und Organisation eines verbundübergreifenden Masterstudiengangs „Data Science“
- Organisation von Weiterbildungen, Schulungen und Lehrformaten für das Thema Data Science und Digitalisierung, insbesondere für anwendungsbezogene und experimentelle Forschende und Studierende

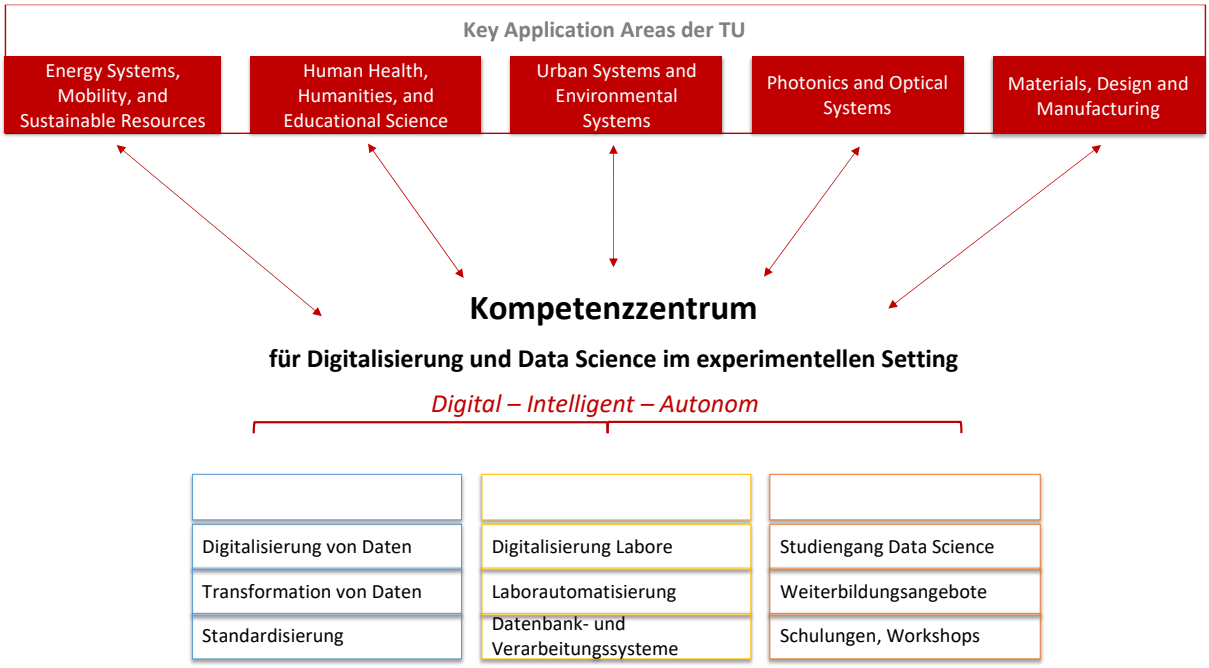

Abbildung 1: Konzeptionelle Einbettung des Kompetenzzentrum in die Key Application Areas der TU Berlin

Das neue Kompetenzzentrum soll einer fächerspezifischen Einbettung zugrunde liegen, denkbar wären hier z. B. die Key Application Areas der TU Berlin (vgl. Abb. 1). Ziel ist aber eine Etablierung innerhalb der Berlin University Alliance um das gesamte Potential des integrierten Standortes zu nutzen. Die TU Berlin übernimmt zunächst für die Pilotphase die Koordination und Federführung. Durch zum Beispiel die Zusammenarbeit im Rahmen des Si-M existiert bereits eine enge Verknüpfung zur Charité – Universitätsmedizin Berlin. Hier soll ein erster Pilot gestartet werden, dessen Ziel die

Bei der Neu- und Weiterentwicklung der Core Facility am Si-M werden Strategien entwickelt und implementiert, die folgende Elemente enthalten: (1) Analyse und Automatisierung von Laborarbeitsabläufen, (2) Entwicklung einer Datenbank in der sowohl die Experiment-spezifischen Daten als auch die gesamten Metadaten in einer Form abgelegt werden, damit auf diese mit verschiedenen Auswerteprogrammen, Simulationsmodulen etc. zugegriffen werden kann. Die Gesamtdaten eines Experiments bzw. Projekts können vollumfänglich auch Partnern zur Verfügung

gestellt werden. (3) Anbindung von Geräten an die Datenbank und Kommunikation untereinander um Workflows transparent zu machen und logisch zu verbinden.

Durch das Pilotprojekt im Rahmen des Si-M soll ein Grundstein gelegt werden für die systematische Erweiterung in weitere Core Facilities. Die finale Struktur wird parallel in dieser Phase gemeinsam mit Beteiligten der Berlin University Alliance erarbeitet. Dazu soll zunächst eine vollständige Landkarte mit möglichen Beteiligten erstellt und diese u. a. im Rahmen von Kurzworkshops zusammengebracht werden. Existierende Projekte, Strukturen, Gruppen und Lehrangebote an den Häusern (vgl. Abschnitt 1) werden dabei als Basis genutzt werden.