SocietyByte
Wissenschaftsmagazin des BFH-Zentrums Digital Society
https://www.societybyte.swiss

Potenziale von Big-Data im Gesundheitswesen

Autoren : Murat Sariyar, Konrad Walser

Datum : 22. Mai 2017

Aufgrund der zunehmenden IT-Unterstützung produzieren Spitäler eine Vielzahl digital vorgehaltener Informationen. Es stellt sich daher immer häufiger die Frage, welchen Nutzen eine Analyse und Verwendung von großen heterogenen Datenmengen (Big-Data) bieten kann.

Eine Definition für Big-Data lautet [1]: „By definition, big data in healthcare refers to electronic health data sets so large and complex that they are difficult (or impossible) to manage with traditional software and/or hardware; nor can they be easily managed with traditional or common data management tools and methods.“

Diese Definition kann der häufig anzutreffenden **3V-Definition** insoweit vorgezogen werden, als sie deutlich macht, dass man zusätzliche Anstrengungen zu vorhandenen Lösungen wie „klassischen“ Data Warehouses unternehmen muss [2]. Die drei V‘s rekurrieren auf **Volume** (Größe), **Variety** (Heterogenität) und **Velocity** (Geschwindigkeit) von Daten. Variety und

Velocity lassen sich dabei dem Begriff „complex“ in der obigen Definition zuordnen, was ein weiteres Problem der 3V-Definition anzeigt, da sich die Komplexität durch ganz andere Eigenschaften als Geschwindigkeit manifestieren kann.

Als Daten für ein Big-Data-System in Spitälern kommen neben den klassischen strukturierten klinischen und verwaltungsbezogenen Daten, die im Krankenhaus vorwiegend im klinischen Informationssystem vorhanden sind, unter anderem folgende Formen vor:

- Omics-Daten (beispielsweise DNA- oder Proteinsequenzdaten)
- Medizinische Referenzdaten aus externen Quellen (beispielsweise Datenbanken zur Pharmakovigilanz oder klinischen Studien)
- Streamdaten von Software auf technischen Geräten (beispielsweise von MRT-Geräten oder mHealth-Apps)
- unstrukturierte Textdaten (beispielweise Arzt- und Pflegeberichte,)
- Umweltdaten (beispielsweise über Ereignisse, Krankheitsentwicklungen und das Wetter).

Die Verknpüfung dieser Daten erfordert syntaktische und semantische Harmonisierungen, bevor diese in Datenbanken zur weiteren Analyse genutzt werden können. In der Medizin sind zur semantischen Harmonisierung in der Vergangenheit u.a. Ontologien und Standards wie UMLS, SNOMED-CT, DICOM, LOINC entwickelt worden. Geht es über die gemeinsame Speicherung hinaus – auch um ein horizontales und vertikales Verknüpfen der Daten – sind Linkage-Verfahren auf Wert-, Record- und Ontologie-Ebene vorzusehen.

Für den Big-Data-Kontext sind neuartige Speicherformen erforderlich:

- Dazu gehören in erster Linie NoSQL-Architekturen, die es erlauben eine hohe Anzahl unterschiedlichster Objekte zu speichern, ohne starre Schemavorgaben machen zu müssen. Vertreter sind beispielsweise ([3], [4]) Cassandra, SAP HANA, CouchDB, MongoDB, ArrangoDB und Virtuoso.
- Neben solchen Datenbanken benötigen Big-Data-Anwendungen gegebenenfalls auch angepasste Dateisysteme mit den Möglichkeiten, mehrere Millionen Dateien effizient und redundant zu speichern,
- aber auch Kompressionsverfahren und erweiterte Data-Warehousing-Funktionalitäten.

Apache Hadoop ist ein Framework, das all diese Bereiche zu adressieren beansprucht [5]. Dazu nutzt es unter anderem das Hadoop Distributed File System (HDFS) als Dateisystem, HBase als Datenbank, MapReduce beziehungsweise gerichtete azyklische Graphen als Prinzip der verteilten Berechnungen und Hive als Data Warehouse [6].

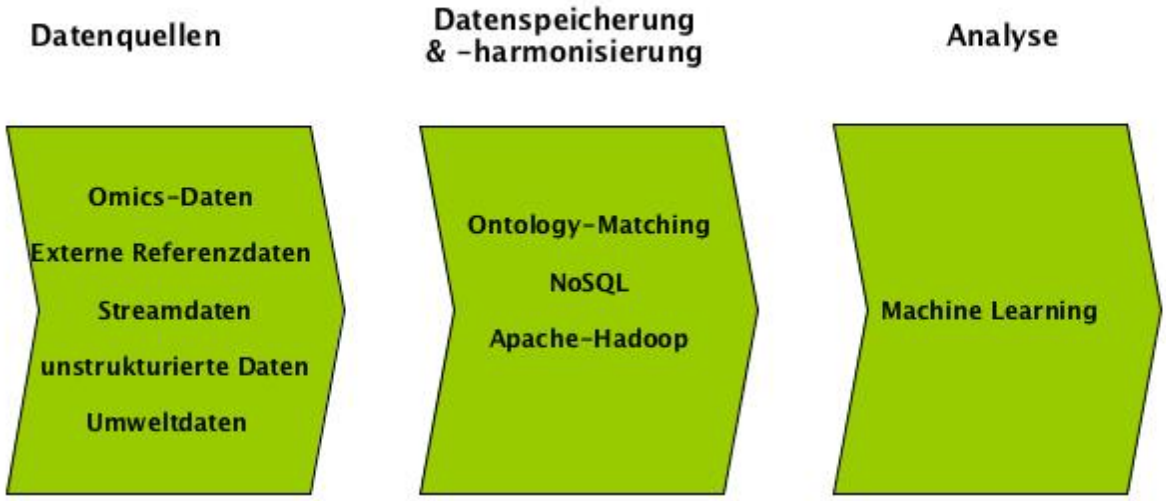

Abbildung 1: Auszeichnende Komponenten von Big-Data in Datenquellen, bei Harmonisierung/Speicherung und Analyse.

Zur neuartigen Analyse solcher Daten eignen sich vor allem Methoden aus dem Bereich des maschinellen Lernens. Bekannte Vertreter sind Support-Vector-Machines, Random-Forests, künstliche neuronale Netze und Conditional-Random-Fields. Eine zusammenfassende Graphik zu den Besonderheiten von Big-Data in den Bereichen Datenquellen, Datenspeicherung und Analyse präsentiert Abbildung 1.

1. Use-Case: Verbesserung der Diagnose von Brustkrebs
Brustkrebs zeichnet sich durch viele unterschiedliche Subtypen aus, die jedoch noch nicht en Detail studiert und in ihren Charakteristika bekannt sind. Wichtige onkologische Daten zu Brustkrebs sind Omics-Daten (Metabolit-, Protein- und genetische Profile), Bilddaten (Computertomographie, Magnetresonanztomographie, etc.) sowie klinische Daten wie Tumor staging und Labordaten. Um zu einer verbesserten Diagnose und damit Therapie zu kommen, sollen alle Datenquellen in eine Big-Data-Diagnose-Pipeline integriert werden und zu einem stabilen Diagnosemechanismus führen. Dabei sind vor allem für die Omics-Daten Fragen nach Real-Time-Analysen, In-Memory-Verarbeitung, Parallelisierung und Größe des benötigten langfristigen Speichers zu klären. Aktuell gibt es noch keine maschinellen Lernenverfahren, die in klinischen Studien validiert worden sind (was für die Anwendung im Behandlungskontext sehr entscheidend ist), insoweit ist für entsprechende Analysen zumeist auf herkömmliche statistische Analysen zu rekurrieren.

2. Use-Case: Aufsetzen einer Datenbank zur onkologischen Forschung
Eine integrierte Big-Data-Anwendung kann zur Beantwortung komplexerer Fragestellungen und Hypothesen im Bereich Onkologie führen. In erster Linie sind forschende Ärzte in der Onkologie daran interessiert, dass möglichst viel Wissen, auch aus externen Datenquellen, in eine Forschungsinfrastruktur einfließt. Zeitkritische Aspekte sind in diesem Fall eher selten.

Interpretationen der Ergebnisse spielen hier jedoch eine größere Rolle, weil es bei onkologischen Fragestellungen um die Erkennung neuartiger Muster geht und nicht um die gleichbleibende Nutzung von Daten in einem komplexen aber bekannten Diagnoseprozess. Das bedeutet konkret, dass die Einbindung von Methodikern und Visualisierungen in diesem Fall sehr wichtig ist. Der Fokus auf neue Hypothesen bedingt auch, dass man eventuell sehr viel mehr Daten und (Zwischen-)Resultate speichert, als dies aus ökonomischen Gründen angebracht erscheinen könnte.

3. Use-Case: Geschäftskennzahlen

Die Geschäftsführung eines Krankenhauses interessiert sich mehr für eine Gesamtstrategie , die auch zu einem ökonomischen Optimum führt, als für einzelne Verbesserungspotenziale.. Eine mögliche Frage könnte etwa lauten: Welcher Mix an DRG's in welchen Mengen führt zu einem ökonomischen Optimum für das Krankenhaus? Zur Beurteilung einer entsprechenden Strategie eines Krankenhauses gehören etwa avisierte Klinikauslastung, Zufriedenheit der Mitarbeiter oder Lieferantenbeziehungen (zu Großhandels-Apotheken, Pharmadistributoren oder etwa Lieferanten von Lebensmitteln, Verbrauchsmateriallieferanten, etc.). Von Interesse kann hier etwa auch die Optimierung der Zuweiserpopulation sein. Typischerweise erfordern solche Analysen lediglich ein klinisches Data Warehouse mit OLAP-Funktionalität (Online Analytical Processing). Insoweit dient dieser Fall der exemplarischen Klarstellung dessen, dass man nicht jede digitale Datennutzung im Dunstkreis von Anstrengungen zu Big-Data unter Big-Data subsumieren sollte.

Ein großer Problembereich von Big-Data betrifft die Qualität und Validität der Ergebnisse. Da große Datenvolumen nicht mehr rein manuell hinsichtlich Plausibilität und Qualität beurteilt werden können, braucht es etablierte und automatisierbare Verfahren, um die Güte der generierten Daten beurteilen zu können. Weitere Probleme betreffen den Datenschutz und ethische Sachverhalte zur Datenverwendung etwa von Personen- und Krankheitsdaten. Je mehr Daten bezüglich der Patienten gesammelt werden, desto größer wird das Risiko, dass Daten missbraucht werden, auch dann, wenn Daten als anonymisiert deklariert werden. Gerade hochdimensionale Daten sind jedoch schwer zu anonymisieren. Dies macht in der Data-Value-Chain weitere organisatorische und technische Absicherungen notwendig.

Beispielhafte Maßnahmen sind: Ausarbeitung klar definierter Policies, Konsens über die Anwendung dieser Policies, Arbeiten mit klaren und sanktionierenden „Terms of Use“, Monitoring aller Aktivitäten in Zusammenhang mit den entsprechenden Daten, Zugangsermöglichung zu den Daten nur über dezidierte dafür vorgesehene Rechner und der Einsatz mehrschichtiger Firewalls.
Ob die Anschaffung und Implementierung von Big-Data-Technologien geboten erscheint, hängt von der adäquaten Einschätzung dessen ab, was man an Ergebnissen erwarten kann und welche Risiken existieren. Die Vernetzung von Leistungserbringern im Gesundheitswesen wird jedoch immer häufiger zu Datensammlungen führen, welche die Big-Data-Nutzung unumgänglich machen.

Literatur

1. Rubin DL, Desser TS. A Data Warehouse for Integrating Radiologic and Pathologic Data. J Am Coll Radiol. März 2008;5(3):210–7.
2. Marz N, Warren J. Big Data: Principles and best practices of scalable realtime data systems. Greenwich: Manning Publications; 2015.
3. Stonebraker M. SQL databases v. NoSQL databases. Commun ACM. 2010;53(4):10.
4. Jing Han, Haihong E, Guan Le, Jian Du. Survey on NoSQL database. In 6th International Conference on Pervasive Computing and Applications (IEEE). 2011;363-366.
5. Nandimath J, Banerjee E, Patil A et al. Big data analysis using Apache Hadoop. In In Proceedings of 14th International Conference on Information Reuse and Intergration (IEEE); 2013;700–703.
6. Thusoo A, Sarma JS, Jain N, Shao Z, Chakka P, Anthony S, u. a. Hive: a warehousing solution over a map-reduce framework. Proc VLDB Endow. 2009;2(2):1626–1629.