



COMMENTARY

Open Access



Is dementia research ready for big data approaches?

Martin Hofmann-Apitius

Abstract

The “big data” paradigm has gained a lot of attention recently, in particular in those areas of biomedicine where we face clear unmet medical needs. Coined as a new paradigm for complex problem solving, big data approaches seem to open promising perspectives in particular for a better understanding of complex diseases such as Alzheimer’s disease and other dementias. In this commentary, we will provide a brief overview on big data principles and the potential they may bring to dementia research, and - most importantly - we will do a reality check in order to provide an answer to the question of whether dementia research is ready for big data approaches.

Keywords: Big data, Data interoperability, Semantic harmonization, Disease modeling, Data mining, Disease mechanisms

Background

According to Wikipedia, “*Big data is a broad term for data sets so large or complex that traditional data processing applications are inadequate. Challenges include analysis, capture, data curation, search, sharing, storage, transfer, visualization, and information privacy. The term often refers simply to the use of predictive analytics or other certain advanced methods to extract value from data, and seldom to a particular size of data set*” [1]. The wide use of the term has contributed to its fuzziness, and the usual IT marketing Newspeak has done its part, too. However, in the original definition by the Gartner group [2] there are three aspects that constitute “big data”; we will use them for the discussion of big data in dementia research. Big data is characterized by:

- **Volume of data.** The “data flood” that we experience in biomedicine, for example, through the enhanced capabilities in the area of genome and transcriptome sequencing technologies, but also in the area of neuroimaging and clinical data
- **Velocity of data.** Usually this “V” is associated with the processing of real-time data analysis (for example, from sensor networks or online trading), but in its

core, this concept also covers the heterogeneity of data with respect to time scales.

- **Variety of data.** This is the most interesting aspect of big data when it comes to biomedicine. Linking heterogeneous data is a key step in big data analytics, and shared semantics for metadata annotation play an important role in data integration.

A fourth aspect has recently been added to the big data concepts; it covers the issue of heterogeneous data quality and the need for curation:

- **Veracity of data.** This concept deals with the need for quality assessment, pre-processing, and curation of data. This is of particular importance in biomedicine, where the quality of data and annotations varies in wide ranges.

Given the complexity of neurodegenerative diseases, it is not surprising that we observe a significant heterogeneity of data in that scientific area. Heterogeneity of data concerns their mode (omics data, imaging data, clinical data → *variety*) as well as their quality (as determined by statistics criteria and with respect to available metadata → *veracity*). Big data analytics makes use of a wide range of data integration, modeling, and mining

Correspondence: martin.hofmann-apitius@scai.fraunhofer.de
Department of Bioinformatics, Fraunhofer Institute for Algorithms and Scientific Computing (SCAI), 53754 Sankt Augustin, Germany

strategies in order to understand and predict systems behavior in complex systems. The expectations on complex problem solving capabilities that come with big data approaches are high in the research community working on dementia. As a first big data challenge in Alzheimer's research, the DREAM challenge has been launched recently [3].

Promising advances, but a reality check is needed

Big data approaches differ substantially from established biostatistics approaches. Whereas biostatisticians try to keep the number of variables low and put great effort into controlling the experiment, big data analytics accepts as a fundamental premise the heterogeneity of data with respect to both quality and type. Big data approaches try to understand complex systems and try to overcome problems with lacking data by a virtuosic combination of data, knowledge, and imputation of missing values. Interoperability of data and knowledge is key to big data approaches and, therefore, shared semantics (for example, controlled vocabularies; ontologies) play a crucial role not only for metadata annotation, but also for data integration and information extraction procedures. In dementia research, we deal with a wide range of data coming from different levels: omics technologies produce large amounts of quantitative data (such as gene expression data) or qualitative data (for example, single nucleotide polymorphism, SNP, data); neuroimaging generates huge amounts of imaging data that require complex image analysis workflows to extract features from images that can be used in integrative modeling and mining approaches. Dementia research is therefore inherently multi-modal (having different modes of data acquisition) and multi-scalar (ranging from the molecular scale (omics) to the organism (neuroimaging; clinical and cognitive data) and population scales (epidemiological data)).

A reality check of the current situation in dementia with respect to the adoption of big data principles leads us to the DREAM challenge, the first data analytics challenge in the area of research on Alzheimer's disease [3]. The tasks in this challenge were highly relevant for current Alzheimer's research, but not too complex and aimed at identifying those features that were informative for the prediction of "cognitive scores 24 months after initial assessment" (subchallenge 1); to "predict the set of cognitively normal individuals whose biomarkers are suggestive of amyloid perturbation" (subchallenge 2); and to "classify individuals into diagnostic groups using MR images" (subchallenge 3).

The data sets made available for the DREAM challenge did not bring along the challenge to harmonize and curate them across different scales or modes of measurement. In that respect, the DREAM challenge on

Alzheimer's disease was not truly a big data challenge. It was rather a statistical quantitative data mining exercise without a decent biological context or knowledge component to integrate. However, attempts to collect and to centrally provision research data in the dementia arena (neuGrid4you [4]; AETIONOMY [5]; Alzheimer's & Dementia knowledge resource [6]) force us to pay much more attention to the most important V's: **V(ariety)** and **V(eracity)**. Over the last three years, our team has spent considerable effort on the quality assessment and curation of all publicly available omics data in the area of neurodegenerative diseases. In order to represent the relevant knowledge in a computable form, we and other groups have generated models of disease [7, 8] that represent a good part of the knowledge about Alzheimer's and Parkinson's diseases and make this knowledge amenable for computer-based reasoning approaches [9]. Together with curated omics data and additional efforts on clinical data, these models form the basis for the most comprehensive knowledge base on neurodegenerative diseases, the AETIONOMY knowledge base [10]. This knowledge base will support future big data approaches in dementia research by harmonized annotations across heterogeneous data sets and tight integration of disease models, literature-based knowledge, and primary experimental data. Other resources that are currently being built include the AMP AD partnership [11]; a dedicated topic on big data in Alzheimer's research can be found on the website indicated in Ref. [12].

Future trends and emerging technologies

A spectrum of currently emerging technologies will add to the **V(olume)** of data relevant for dementia research: next-generation sequencing (NGS) and in particular RNAseq technologies will generate high-quality gene expression data; epigenetics studies in the dementia context will add an entire level of new information. Patient-specific iPS cells will be analyzed simultaneously for gene expression, epigenetics mechanisms, and proteomics (including pathway regulation). More data will also be generated at the clinical level: large observational studies (such as the Rhineland Study conducted by the German National Dementia Research Center [13]) will produce huge amounts of clinical and imaging data. The interpretation of data generated at various different levels and including data that come from either cell and tissue cultures or animal models will require new approaches for integrative data analysis. Big data principles will therefore become even more relevant for integrative modeling and mining and any sort of "systems approach" in dementia research. The computable Alzheimer's disease model shown in Fig. 1 (a model representing cause-and-effect relationships under disease conditions) captures and represents knowledge about disease processes in Alzheimer's

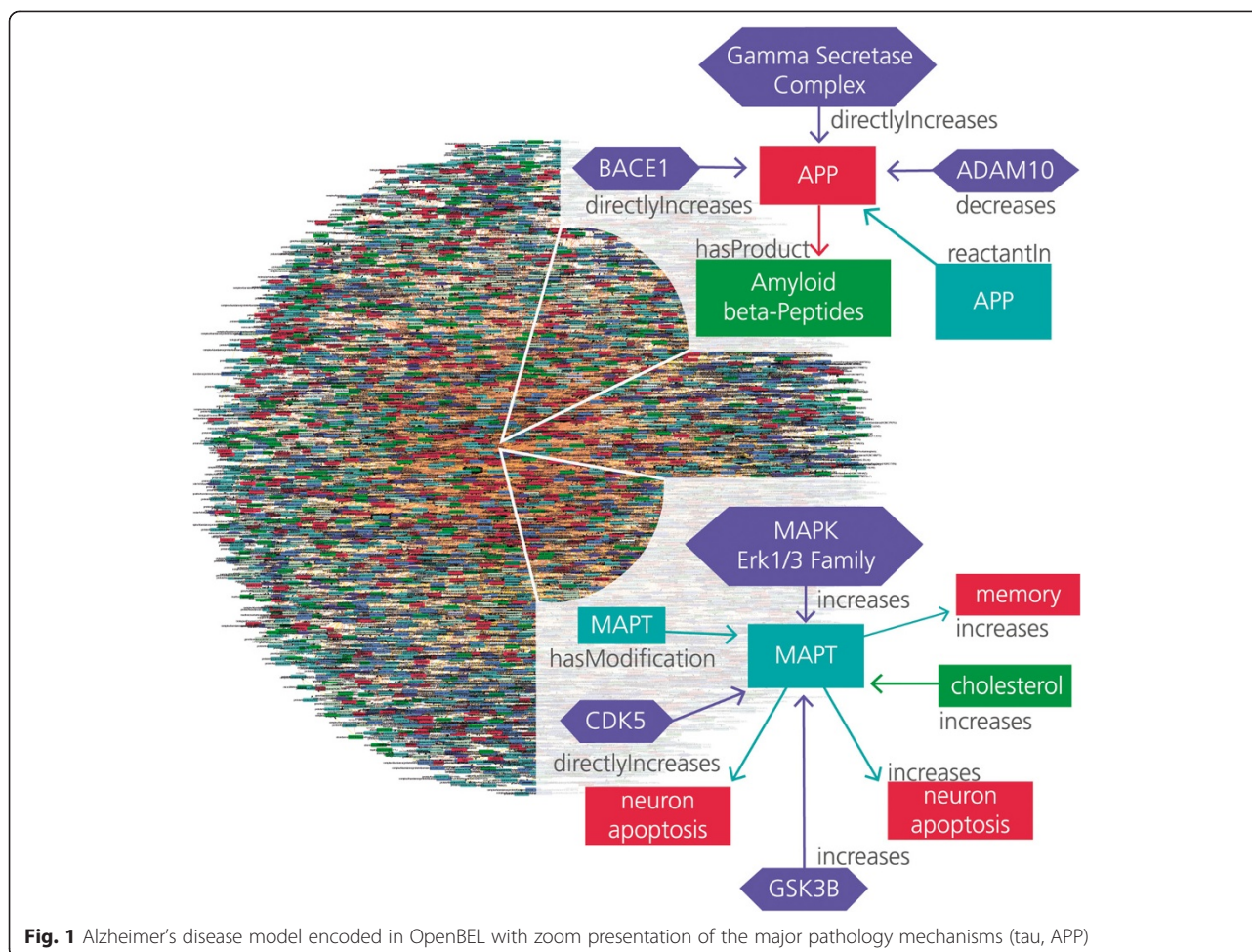


Fig. 1 Alzheimer's disease model encoded in OpenBEL with zoom presentation of the major pathology mechanisms (tau, APP)

disease and is ideally suited to support the enhanced interpretation of big data in the Alzheimer's research area.

Conclusions

We can expect that, in the near future, data production at all levels - from the omics level to the clinical and population level - will increase with the same rate in dementia research that we can observe in other indication areas. The need for increased interoperability of data (and knowledge!) will simultaneously increase, and substantial effort will be required to cope not only with the rapid growth of data volume, but also with the notorious lack of interoperability of data, information, and knowledge. We will see more ambitious mining scenarios in big data challenges in the future, and there is no doubt: integrative modeling and mining approaches will come, and they will have a strong impact on dementia research.

Competing interests

The author declares that he has no competing interests.

Author's information

Martin Hofmann-Apitius is Professor of Applied Life Science Informatics, University of Bonn, Germany and Head of the Department of Bioinformatics, Fraunhofer Institute for Algorithms and Scientific Computing (SCAI), Sankt Augustin, Germany.

Received: 12 May 2015 Accepted: 13 May 2015

Published online: 22 June 2015

References

1. Wikipedia entry on big data. http://en.wikipedia.org/wiki/Big_data. Accessed 4 May 2015.
2. Gartner Group IT glossary. <http://www.gartner.com/it-glossary/big-data>. Accessed 4 May 2015.
3. The Alzheimers Disease Big Data DREAM Challenge 1 website. <https://www.synapse.org/#!Synapse:syn2290704>. Accessed 4 May 2015.
4. N4U - neuGRID for you expansion of neuGRID services and outreach to new user communities. <https://neugrid4you.eu/>. Accessed 4 May 2015.
5. Developing a "mechanism-based taxonomy" of Alzheimer's and Parkinson's Disease. <http://www.aetionomy.eu/>. Accessed 4 May 2015.
6. Alzheimer's Today The continuing symposium on Alzheimer's and dementia research, care, policy and education. <http://adjresourcenter.alzdem.com/>. Accessed 4 May 2015.
7. Kodamullil AT, Younesi E, Naz M, Bagewadi S, Hofmann-Apitius M. Computable cause-and-effect models of healthy and Alzheimer's disease states and their mechanistic differential analysis. *Alzheimers Dement*. 2015. doi:10.1016/j.jalz.2015.02.006.

8. Fujita KA, Ostaszewski M, Matsuoka Y, Ghosh S, Glaab E, Trefois C, et al. Integrating pathways of Parkinson's disease in a molecular interaction map. *Mol Neurobiol.* 2014;49:88–102.
9. Catlett NL, Bargnesi AJ, Ungerer S, Seagaran T, Ladd W, Elliston KO, et al. Reverse causal reasoning: applying qualitative causal knowledge to the interpretation of high-throughput data. *BMC Bioinformatics.* 2013;14:340.
10. Access to the AETIONOMY knowledge base with curated and re-annotated data (to access the AETIONOMY knowledge base you will have to get registered first). <http://aetionomy.scai.fhg.de/>.
11. Accelerating Medicines Partnership (AMP) Alzheimer's Disease Knowledge Portal. doi:10.7303/syn2580853 <https://www.synapse.org/#!Synapse:syn2580853>. Accessed 4 May 2015.
12. Alzheimer's Today The continuing symposium on Alzheimer's and dementia research, care, policy and education – Big Data featured posts. <http://adjresourcenter.alzdem.com/BigData>. Accessed 4 May 2015.
13. Population health sciences in the German National Center for Dementia Research (DZNE). Population Studies. <http://www.dzne.de/en/research/research-areas/population-studies.html> Accessed 4 May 2015.

**Submit your next manuscript to BioMed Central
and take full advantage of:**

- Convenient online submission
- Thorough peer review
- No space constraints or color figure charges
- Immediate publication on acceptance
- Inclusion in PubMed, CAS, Scopus and Google Scholar
- Research which is freely available for redistribution

Submit your manuscript at
www.biomedcentral.com/submit

