Buchtitel: „‘Forensigraphie‘ – Möglichkeiten und Grenzen IT-gestützter klinisch-forensischer Bildgebung“

- **Titel des Beitrags** : Big data in medical imaging, forensics and beyond

- **Name und Kurzbeschreibung der Autorin/des Autors** : Henning Müller, HES-SO

Henning Müller studied medical informatics at the University of Heidelberg, Germany, then worked at Daimler-Benz research in Portland, OR, USA. From 1998-2002 he worked on his PhD degree at the University of Geneva, Switzerland with a research stay at Monash University, Melbourne, Australia in 2001. Since 2002 Henning has been working in medical informatics at the University Hospitals of Geneva where he habilitated in 2008 and was named titular professor in 2014. Since 2007 he has been a professor in business informatics at the HES-SO Valais in Sierre and since 2011 he has been responsible for the eHealth unit in Sierre. Henning was coordinator of the Khresmoi project, scientific coordinator of the VISCERAL project, initiator of the ImageCLEF benchmark. He has authored over 400 scientific papers, is in the editorial board of several journals and reviews for many journals and funding agencies around the world.
For 2015-2016 Henning was a visiting professor at the Martinos Center in Boston, MA, USA part of Harvard Medical School and the Massachusetts General Hospital (MGH) working on collaborative projects in medical imaging and system evaluation among others in the context of the Quantitative Imaging Network of the National Cancer Institutes.

- **Abstract**

Motivation: Big data and deep learning are currently hype topics that are covered in large science journals and by companies alike to potentially solve many of our future problems, from self-driving cars to home robots and intelligent medicine. Beyond the hype are many techniques that are very interesting and can include medical images as well as data from forensics that encompasses varied sources combined, from images, free text, sematic terms to structured data. The potential for useful applications that allow searching large amounts of data of varying sources and finding the needle in the haystack can help many scientific areas and in routine applications.

Objectives: The objective of this article is to give an overview of the fields of big data and deep learning in a context of medical image analysis and forensics. For this, the main concepts are explained and a few example projects that the author is involved in are described in more detail.

Results: This article gives an overview of several scientific projects in the areas of sharing data sets and data availability for medical imaging and related fields, as the data sets are the main foundation for the applications that can then exploit data and knowledge about the data. Evaluation infrastructures for comparing tools on large scale data will then be explained, as it is important to compare algorithms on realistic scenarios and focus theoretical scientists onto concrete and real problems. A few examples for big data applications and techniques used in this context such as deep learning are explained next. The focus is on the medical imaging field in a more general sense, but much of this can easily be extended to forensics or any other related field that is rich in images and associated meta-data that are collected systematically.

Conclusions: The creation of large digital data sets that contain images and text or structured meta-data has become much easier with costs for image creation and storage being strongly reduced

every year. As also new infrastructures (i.e. using cloud computing) allow evaluation of several algorithms on restricted data sets (also called crowdsourcing of algorithm development), the possibility to run tools on extremely large data sets is given in the biomedical and related fields. Tools for data analysis exist and only need to be adapted. Several companies have shown the potential value of such applications, and it will likely continue to influence the way we manage and exploit data sets, particularly when including images or videos.

- **Bestichwortung des Beitrags für das Register**

Big data, Image retrieval, machine learning, artificial intelligence, multimodal data analysis

- **Beitragstext in etwaiger Gliederung**

1. Introduction

The term **big data** has become frequently used and abused over the last ten years in a variety of contexts [1]. Many companies use big data sets, for example the web giants Facebook and Google, to optimize the placement of advertisements on the web pages that their users access. With the pay-par-click model, these companies only get paid if an advertisement link is actually clicked by one of the users, so the right placements of the advertisements for a user is essential. This can include much knowledge on the users that both Google and Facebook have for many of their clients (via a variety of applications such as Gmail, search history, use of mobiles applications or Facebook with Likes and other contacts or posted contend). Thus, both companies use big data analysis to optimize their revenue and are very successful with it. Many examples for such use of data to optimize a company's operation include for example a barbecue restaurant [2], where equally data is used to optimize various aspects of revenue and user contact. Customer satisfaction can be measured and new products developed, stock reduced and trends found much earlier than without using real time data analysis.

The medical field has been an obvious extension to big data analysis as the potential is enormous with massive amounts of data being produced (increasingly in digital format that allows exploitation); and the part of medical care in the GDP has been rising in most industrial countries in the past years. Some articles have predicted that IBM's Watson computer may soon be better than physicians [3] but to the best of my knowledge there is no detailed evaluation of the quality of Watson available to date that compares it with other techniques. Similarly, for medical imaging in [4] it is mentioned that such data occupy around 30% of world storage, thus much more than the big video platforms such as YouTube, that equally treat massive amounts of data. In difference to text or structured data, images are more complex to use, because meaningful visual features (signatures or visual characteristics, such as colors, shapes and texture) need to be extracted and the image pixels themselves are not directly meaningful information. Ways to exploit image data are explained in more details below.

To start with what is generally understood by the term "big data" several definitions exist. This is not just about the volume of the data but also its complexity. As storage capacities grow, so do the amounts that are considered "big data". In general, a data set can be considered big data if "it is too big or complex to be treated with traditional means". Usually when we speak about **big data** three characteristics are most frequently mentioned, **the three "V"s**:

- large **Volume** (from Terabytes to Petabytes and Exabytes);
- large **Velocity** (new data arriving continuously requiring updating of the analysis);
- large **Variety** (text, imaging, signals, different types of data and in part unstructured);

Thus, a problem is considered a big data problem only when all three occur in a data set, so the volume is large, there are new data arriving and the data are no simple, structured data but contain a variety of data types and also unstructured data.

An additional "V" often mentioned is also **Veracity,** as in very large data sources we might have data quality differences or even contradicting data sources that require to be taken into account. This holds particularly true in the medical field where data entry in patient records often happens under time pressure or in stressful situations, and the reuse of the data is most often not considered at the moment of data entry. Even the automatically generated DICOM (Digital Imaging and Communications in Medicine) headers of medical images contain fields that have up to 30% errors [5]. A fifth "V" can be the **Value** of the data, as some data might have more value that other parts and this can be taken into account for an analysis.

As big data does not give any differences between the exact content that is available there is also the separation between **thin data** and **thick data**. Many used data sets such as web clicks, logs of search terms in web search engines or the analysis of a shopping baskets are **thin data**, as little data in a single event are available. This often allows to quantify a problem or phenomenon, but it often does not allow for understanding the reasons for it. Market basket analysis (analysis of what people buy in supermarkets) allows to analyze what people buy together and how much they buy but rarely helps to understand why specific changes appear over time. **Thick data** on the other hand aims at finding much data on fewer events, for example social science studies that follow shoppers and have detailed interviews with them to understand the reasons for specific behavior. The two approaches can usually be applied together for analyzing very different aspects of a same problem.

When mentioning medical big data, the first area that comes to mind is most often **genomics**, as the human genome contains a large amount of data (3.2 billion base pairs, or 3.2 GB for a single genome, but this can easily be compressed further) for a single patient, even though generally only a small part of it may be relevant in a specific situation. Getting larger data sets is not easy because data can not easily be shared, as they are confidential. To interpret genomic data, links with the patient history and disease patterns usually need to be made, thus corresponding to the various aspects of big data (volume, variety, velocity and veracity).

In terms of the quantity of data production **medical imaging** is definitely still the largest data producer in medical institutions [4,8,16]; for example, the University hospitals of Geneva produced over 300'000 images per day in 2015. Many data sets have been made available by the NIH (National Institutes of Health), for example via the TCIA[1] (The Cancer Imaging Archive) and many national funding agencies also push for sharing image data sets. Still, region-based annotations and meta-data are not always available, limiting sometimes the usefulness of data or if there is no clear evaluation scenario data can also be used in many different ways, making it effectively hard to compare any two approaches that use the same data.

The **quantified self** and personal big data are domains that have been increasing in importance strongly. Private persons using such devices can control a large number of parameters from heart rate, temperature to numbers of steps walked pretty much 24 hours per day and again create large amounts of sensor data. This is currently often not linked with patient records but the use of the data by physicians can be beneficial, even though there is much discussion about the quality of the sensors in very inexpensive devices [9]. Medical social networks such as PatientsLikeMe[2] also try to collect these data and manually entered forms about mainly patients with rare diseases to subsequently use the data and learn from them. Having regular updated of these data done by patients can create much more complete records over time that allow judging a patient condition better than when only having infrequent encounters with a physician.

Deep Learning [6,7] is currently maybe the hottest topic in artificial intelligence. Its emergence and use was impressive particularly on unstructured data such as images (for example in the ImageNet

[1] http://www.cancerimagingarchive.net/

[2] http://www.patientslikeme.com/

challenge [10]) and other complex tasks such as playing the Go game [14], where the Deepmind[3] system won against one of the best human players. Main drivers for improving deep learning in the past ten years were the use of GPUs that allowed to test many more configurations and architectures of networks in a limited amount of time and thus allowing more complex and deeper structures of the networks.

Also **forensics** has seen many efforts to systematically collect and annotate data from the past to learn for the future [11,12]. This includes several imaging modalities as well as structured data and free text. Even 3D simulations can be used to try reconstructing events that might have led to specific situations. Many of the techniques used for medical image analysis and related multimodal data can likely also be used on data sets in forensics. Recording data from past cases can also constitute a valuable basis for reconstructing new events, but only if the data are acquired properly and can effectively be searched and analysed.

2. Evaluation infrastructures and data sets

This section describes the basis of big data analysis and the tasks related to it. The sharing and availability of the data sets is one major challenge that needs to be addressed. Even though funding bodies see the high potential in sharing data (particularly with expensive manual annotations), they also see the ethics concerns that are associated with sharing data sets. For this reason, the data sets are here in the same section as the evaluation infrastructures, as such infrastructures can also help sharing the data for analysis without actually having to make the data available [13].

2.1 Data sets for big learning and data annotation

As mentioned beforehand, the availability of data is of major importance. Whereas general data such as text and images are available on the Internet it is not always clear in which way the data can be used. Some content providers (i.e. Flickr) attach licenses to content, such as CreativeCommons[4], and thus create a stable framework for data reuse. In the medical field data confidentiality and privacy preservation are extremely important. Thus, informed consent is usually necessary for data use, also when data are anonymized before their use in a research project. As images can contain names or birth dates in the pixel data and also high resolution scans of faces can allow identification, the anonymization is not always easy and most often needs to be manually controlled. Also free text reports can be automatically anonymized but there remains a small risks of revealing potentially identifying information, particularly as the data sets to be shared become extremely large. This creates very restrictive policies for medical data use even though the NIH pushes for more data sharing in all its funded projects. Many funding bodies also favor to make data sets and tasks available in the form of scientific challenges to maximize data use and impact.

In general difficulties in sharing data relate to one or several of the following:

- **Very large** data sets (in the order of several Terabytes) can not be downloaded anymore and also sending hard disks is cumbersome when really large data sets are concerned.
- **Confidential** data can often not be shared among several partners (in the medical field but also for enterprise search or in the investigative domain).
- When creating static data sets time passes between the moment a resource is created and when it is used and thus in domains that **change quickly** (for example in mobile phone providers) this common paradigm is hard to use as work should be done online on the latest data.

2.2 Evaluation infrastructures for large data sets and confidential data

[3] http://www.deepmind.com/

[4] https://creativecommons.org/

Systematic evaluation and measurements have been the basis for many scientific advances and the phase “If you can not measure it, you can not improve it.” (Lord Kelvin) underlines this fact. In many disciplines such as information retrieval systematic comparisons of techniques and comparisons to baselines are fairly old. TREC[5] (Text Retrieval Conference) has started in 1992 to create a workshop where a large variety of techniques were compared on a yearly basis, making data sets available and the tasks and topics and then collecting the results of research systems and discussing the outcomes at a workshop to optimize the tools for future challenges.

ImageCLEF [22] was in 2003 one of the first benchmarks focusing on image retrieval, so finding images with text but also with visual examples, so searching for visually similar images. This allowed comparing many tools and techniques over the years and also to discover the difficulties in organizing such challenges and share data.

In machine learning several platforms have started to operate for doing such Crowdsourcing tasks [21]. Any organization with data and a challenge of analyzing the data or predicting actions (such as Amazon who would like to know which book a person is likely to buy next) can thus create a task on platforms such as Kaggle[6] or TopCoder[7]. Many tasks have price money for the best solution and many research groups in machine learning then participate trying to obtain good results. The organization can thus get for a limited amount of funding a detailed comparison of many techniques that would be impossible to obtain without a large research department.

Based on the challenges of distributing data, quickly changing data and confidential data the VISCERAL project [18,19] created a totally different infrastructure for the evaluation of medical image analysis tools, as can be seen in Figure 1. The idea is that data are not distributed to participants of the challenge but the participants are given access to the data via a cloud computing infrastructure, in the case of VISCERAL using the Microsoft Azure cloud. All data remain in the central cloud storage and participants get a virtual machine to access the small set of training data to make sure that there software works correctly on the data. The test set never needs to get accessed and the challenge organizers then take control over the virtual machines in the test phase and run the algorithms on the full test data set. This infrastructure moves the algorithms to the data and not the data to the algorithms, which seem more practical in the era of big data.

[5] http://trec.nist.gov/

[6] http://www.kaggle.com/

[7] http://www.topcoder.com/

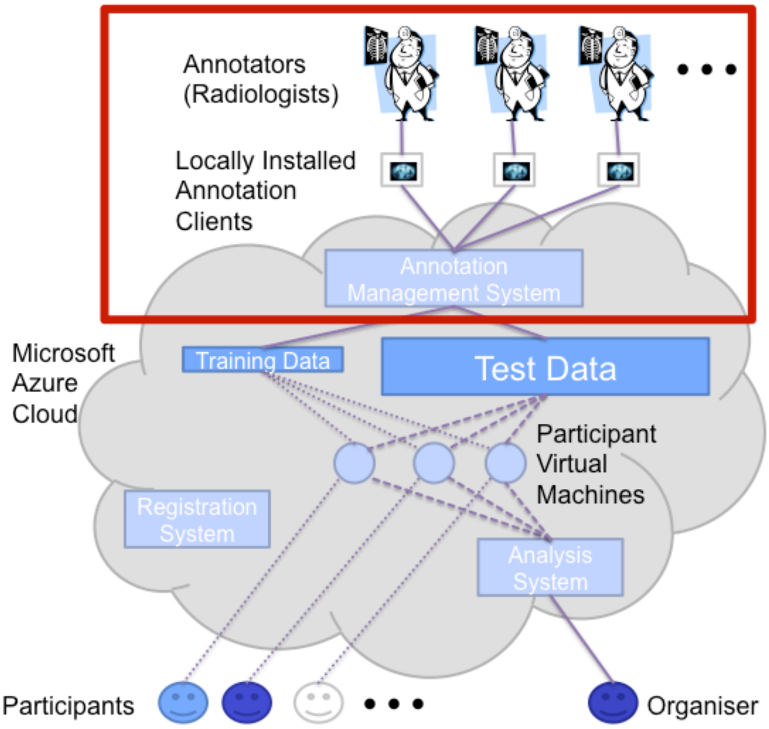


Figure 1: Overview of an Evaluation-as-a-Service infrastructure, where data sets are made available for scientific competitions inside a cloud infrastructure, but only training data are visible and test data are fully protected, thus allowing to use confidential data and avoiding any use of test data to optimize results [17,18].

The model thus has several advantages: no data need to be distributed and thus data are not duplicated but stored only once, so data breaches are not possible. The test data are kept totally confidential and only the algorithms see the data, not the researchers. The test data can always be the latest version of the data. As no data are distributed it is not possible to match data from one set with other sources, which is on of the risks that can lead to discovering the identity of a person. As the data and the executable are available and can be kept long term, the process is also fully reproducible and thus can avoid wrong interpretation of data or even manipulating data. This responds also to [24] that explains why at the moment most published research findings can be considered false.

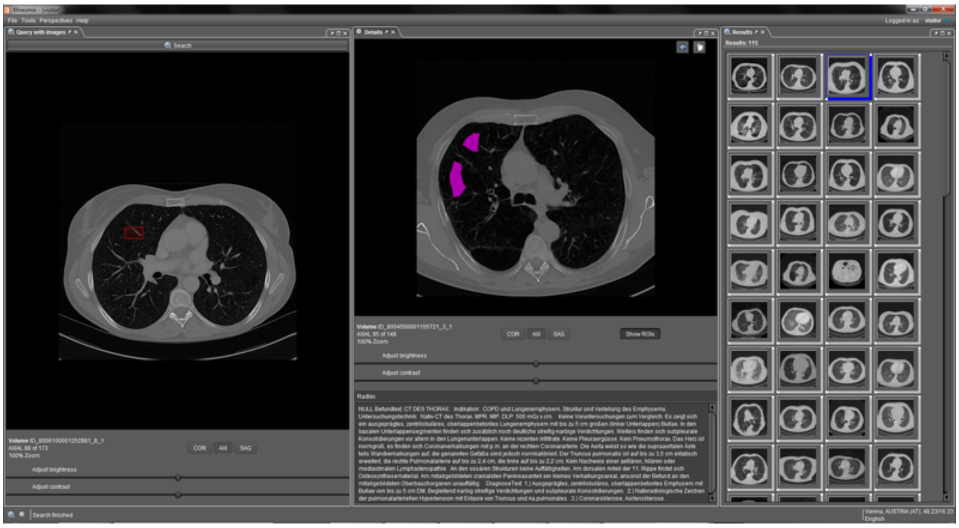
This setup also led to the NCI (National Cancer Institute) to run challenges in a similar format in the Coding4Cancer[8] initiative that is based on cloud-based evaluation [20]. Currently a challenge on mammography data has started and an initiative on lung data is planned.

3. Information retrieval, image analysis and machine learning

The actual tools to make extremely large data sets accessible and searchable rely on a variety of underlying techniques**. Information indexing and retrieval** techniques are like the most visible tools that are used by everyone on a daily basis, for example with the Google or Bing search engines, and also everywhere with related mobile applications. Internet search engines changed our behavior, as we do not need to know everything anymore but it is sufficient to know how to find things. Still, in radiology, text books are more commonly used to find similar patterns to a case that is difficult to diagnose. The search for visual data such as images is a challenge because images do not contain words that can be used directly but important patterns need to be extracted from the pixel data with generic features. Such visual retrieval is also called **content-based image retrieval** [16] and much research has been done on it in the medical domain.

[8] http://www.coding4cancer.org/

The Khresmoi research project [23] has developed several prototypes for image search, for example the prototype seen in Figure 2. Here, a radiologist can mark a small image region and then search for other cases that contain similar image regions. This allows searching in many different ways that are complementary to text. It allows browsing in large data sets that can be without annotations. In many domains search by visual means or by similarity, which offers possibilities also for forensics.



Results: 115

Volume ID_8006100001252881_8_1
AXIAL 88 of 173
100% Zoom

Volume ID_8004500001155721_3_1
AXIAL 95 of 148
100% Zoom

Radlex:

NULL Befundtext: CT DES THORAX: Indikation: COPD und Lungenemphysem. Struktur und Verteilung des Emphysems. Untersuchungstechnik: Nativ-CT des Thorax. MPR, MIP. DLP: 500 mGy x cm. Keine Voruntersuchungen zum Vergleich. Es zeigt sich ein ausgeprägtes, zentrilobuläres, oberlappenbetontes Lungenemphysem mit bis zu 5 cm großen (linker Unterlappen) Bullae. In den basalen Unterlappensegmenten finden sich zusätzlich noch deutliche streifig-narbige Verdichtungen. Weiters finden sich subpleurale Konsolidierungen vor allem in den Lungenunterlappen. Keine rezenten Infiltrate. Keine Pleuraergüsse. Kein Pneumothorax. Das Herz ist normgroß, es finden sich Coronarverkalkungen mit p.m. an der rechten Coronararterie. Die Aorta weist so wie die supraaortalen Äste teils Wandverkalkungen auf, die genannten Gefäße sind jedoch normkalibriert. Der Truncus pulmonalis ist auf bis zu 3,5 cm ektatisch erweitert, die rechte Pulmonalarterie auf bis zu 2,4 cm, die linke auf bis zu 2,2 cm. Kein Nachweis einer axillären, hilären oder mediastinalen Lymphadenopathie. An den ossären Strukturen keine Auffälligkeiten. Am dorsalen Anteil der 11. Rippe findet sich Osteosynthesematerial. Am mitabgebildeten cranialsten Pankreasanteil ein kleines Verkalkungsareal, ansonst der Befund an den mitabgebildeten Oberbauchorganen unauffällig. DiagnoseText: 1.) Ausgeprägtes, zentrilobuläres, oberlappenbetontes Emphysem mit Bullae von bis zu 5 cm DM. Begleitend narbig streifige Verdichtungen und subpleurale Konsolidierungen. 2.) Nativradiologische Zeichen der pulmonalarteriellen Hypertension mit Ektasie von Truncus und Aa.pulmonales. 3.) Coronarsklerose, Aortensklerose.

Figure 2: Interface of a visual search system that allows to mark an image region and then search for other cases and images that contain visually similar regions [23].

Image analysis is necessary to make large amounts of visual data (images, videos, 3D volumes, etc.) available for analysis and also search. The tools for automatically analysis images are often not perfect because there is a large variety in image acquisition conditions, views and content available. Humans easily analyze such images but for a machine this is a difficult task. Each algorithm usually has systematic mistakes. Learning based on human annotations of data can improve the quality of algorithms but human annotations are often expensive particularly when medical experts are required. In the VISCERAL project [18] health experts annotated organs in 3D tomographic data to train and evaluate many automatic algorithms for organ segmentation. By having several algorithms segment unknown cases, the labels of several algorithms can be fused, leading generally to better results than any single algorithm as show in Figure 3.

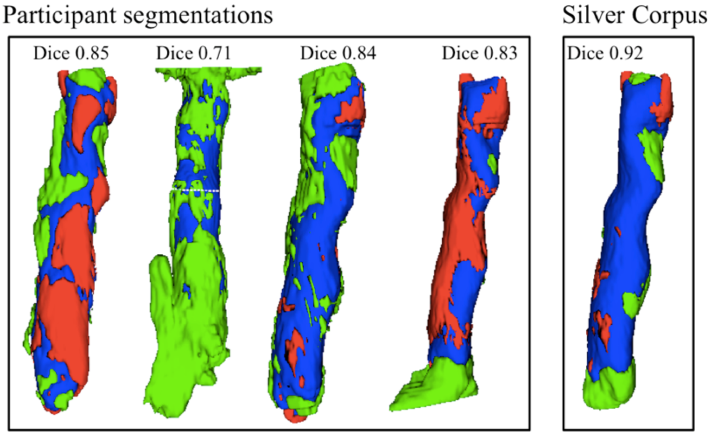


Figure 3: The availability of several algorithms usually allows to reach much better results than any single algorithm if really different algorithms are used. This allows to create silver corpora that are of very good quality and limit the large amount of work for manual annotations [15].

This way of combining labels is often also called label fusion and it allows to create new annotations automatically, maybe not as good as a gold standard but rather a silver standard [15]. These data can then again be used to retrain the machine learning algorithms and potentially improve their outcomes again, creating potentially a loop to improve the outcomes as soon as there are massive amounts of data available for analysis and training and of course computing infrastructure that support this.
Many techniques of machine learning have been developed over the years, from Support Vector Machines (SVMs) that create separation lines between classes of items, random forests that can improve generalization. In the past few years deep learning (deep neural networks) has been a hype topic as the available computing power and training data have led to extremely good results, for example in the ImageNet challenge [25]. As other applications have also had extremely good results based on large computing power and available training data deep learning is now used in basically any application. Many tools are available free of charge such as TensorFlow[9], Caffe[10] or Torch[11]. Many new networks and approaches are frequently developed and the current opportunities seem almost limitless. On the other hand, deep learning basically is a black box, so often the outcome is very good but it can not always be explained. This led to tools visualizing output of layers of the network that can potentially explain results, as for example can be seen in Figure 4 for basic objects that can be detected with deep learning.

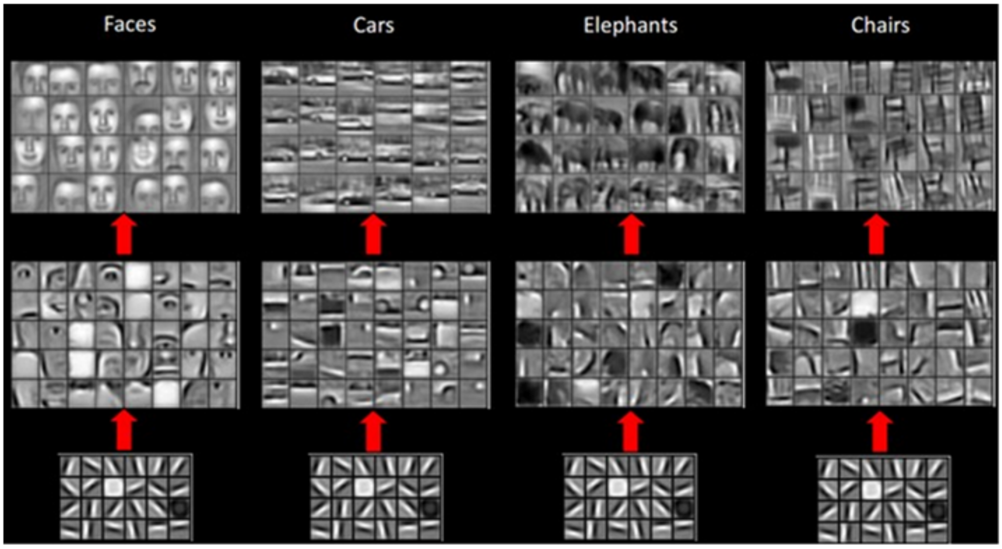


Figure 4: Visualizing of layers in deep neural networks that are linked to detecting specific objects (image taken from http://stats.stackexchange.com/questions/146413/why-convolutional-neural-networks-belong-to-deep-learning).

Thus, the advent of large data sets, computing power and machine learning will likely help with many other problems and ideas in the future.

4. Conclusions

The age of big data is clearly still at its beginning. Many tools and techniques exist and are being used in a variety of settings. Particularly the exploitation of health related data can change the current course of medicine and in related fields. Knowing machine learning can clearly help to organize data and make it accessible. Closed infrastructure do not need to share data but rather executable software that than runs on confidential data sets.

[9] https://www.tensorflow.org/
[10] http://caffe.berkeleyvision.org/
[11] http://torch.ch/

Clearly, there are areas where much work is needed, for example how to make complex machine learning algorithms easy to use and apply. Managing the underlying infrastructure has already started to become invisible with cloud computing that gives access to almost unlimited computing. Visualizing results of search and exploration of large data sets is another are where much work is still necessary even though this area has already seen many new approaches, for example in interactive visualizations and visual analytics.

Big data will likely follow us for the years to come and it will be interesting to see application in many new fields where data are currently being acquire and becoming available in digital form.

5. References

[1] James Manyika, Michael Chui, Brad Brown, Jacques Bughin, Richard Dobbs, Charles Roxburgh, Angela Hung Byers, Big data: The next frontier for innovation, competition, and productivity, Report McKinsey Global Institute, 2011.

[2] Bernard Marr, Big Data At Dickey's Barbecue Pit: How Analytics Drives Restaurant Performance, Forbes, June 2, 2015.

[3] Lauren F Friedman, IBM's Watson Supercomputer May Soon Be The Best Doctor In The World Business Insider, Apr. 22, 2014.

[4] High Level Expert Group on Scientific Data, How Europe can gain from the rising tide of scientific data – final report of the high level expert group on scientific data. http://cordis.europa.eu/

[5] Mark O. Gueld ; Michael Kohnen ; Daniel Keysers ; Henning Schubert ; Berthold B. Wein ; Joerg Bredno ; Thomas M. Lehmann, Quality of DICOM header information for image categorization, Proceedings of the SPIE 4685, Medical Imaging 2002: PACS and Integrated Medical Information Systems: Design and Evaluation, San Diego, CA, USA, 2002.

[6] Yann LeCun, Yoshua Bengio, Geoffrey Hinton, Deep learning, Nature 521 (7553), pages 436-444, 2015.

[7] Jürgen Schmidhuber, Deep learning in neural networks: An overview, Neural Networks 61, pages 85-117, 2015.

[8] Henning Müller, Allan Hanbury, Forschungsanwendungen in der digitalen Radiologie: „Big data“ und Co., Der Radiologe, 2016.

[9] For Wearables, Accurate Sensing Is Tricky, MIT Technology Review. Retrieved July 6, 2016 from http://www.technologyreview.com/review/538416/the-struggle-for-accurate-measurements-on-your-wrist/

[10] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, Alexander C. Berg, Li Fei-Fei, ImageNet Large Scale Visual Recognition Challenge, International Journal of Computer Vision, Volume 115, Issue 3, pages 211–252, 2015.

[11] Barry Dalya, Samir Abbouda, Zabiullah Alib, Clint Slikera, David Fowler Comparison of whole-body post mortem 3D CT and autopsy evaluation in accidental blunt force traumatic death using the abbreviated injury scale classification, Forensic Science International, Volume 225, Issues 1–3, 10 February 2013, Pages 20–26

[12] Klaus Poulsen, Jørn Simonsen, Computed tomography as routine in connection with medico-legal autopsies, Forensic Science International, Volume 171, Issues 2–3, Pages 190–197, 2007

[13] Hanbury A, Müller H, Langs G, Weber MA, Menze BH, Salas Fernandez T, Bringing the algorithms to the data: cloud–based benchmarking for medical image analysis. In: Catarci (Hrsg). Information access evaluation. Multilinguality, multimodality, and visual analytics. Springer, Heidelberg, pages 24–29, 2012

[14] David Silver, Aja Huang, Chris J Maddison, Arthur Guez, Laurent Sifre, George Van Den Driessche, Julian Schrittwieser, Ioannis Antonoglou, Veda Panneershelvam, Marc Lanctot, Sander Dieleman, Dominik Grewe, John Nham, Nal Kalchbrenner, Ilya Sutskever, Timothy Lillicrap,

Madeleine Leach, Koray Kavukcuoglu, Thore Graepel, Demis Hassabis, Mastering the game of Go with deep neural networks and tree search, Nature, Volume 529, pages 484-489, 2016.

[15] Krenn M, Dorfer M, Oscar Alfonso Jimenez del Toro, Henning Müller, Bjoern Menze, Marc-Andre Weber, Allan Hanbury, Georg Langs, Creating a Large-Scale Silver Corpus from Multiple Algorithmic Segmentations, MICCAI medical computer vision workshop at MICCAI. In: MICCAI medical computer vision workshop at MICCAI, Munich, Germany, 2015.

[16] Henning Müller, Nicolas Michoux, David Bandon, Antoine Geissbuhler, A review of content-based image retrieval systems in medicine – clinical benefits and future directions, International Journal of Medical Informatics volume 73, pages 1-23, 2004.

[17] Allan Hanbury, Henning Müller, Krisztian Balog, Torben Brodt, Gordon V. Cormack, Ivan Eggel, Tim Gollub, Frank Hopfgartner, Jayashree Kalpathy-Cramer, Noriko Kando, Anastasia Krithara, Jimmy Lin, Simon Mercer, Martin Potthast, Evaluation-as-a-Service: Overview and Outlook, ArXiv, 2015.

[18] Georg Langs, Henning Müller, Bjoern Menze, Allan Hanbury, Visceral: Towards large data in medical imaging – challenges and directions. In: Medical Content-Based Retrieval for Clinical Decision Support. Springer, Heidelberg, 2013.

[19] Oscar Alfonso Jiménez del Toro, Henning Müller, Abdel Aziz Taha, Katharina Gruenberg, Markus Krenn, Marianne Winterstein, Orcun Goksel, Bjoern Menze, Georg Langs, Marc–André Weber, Tomàs Salas Fernandez, Antonio Foncubierta–Rodríguez, Ivan Eggel, Roger Schaer, András Jakab, Georgios Kontokotsios, Mohammad A. Dabbah, Yashin Dicente Cid, Tobias Gass, Mattias Heinrich, Fucang Jia, Fredrik Kahl, Razmig Kechichian, Dominic Mai, Assaf B. Spanier, Graham Vincent, Chunliang Wang, Allan Hanbury, Cloud–based Evaluation of Organ Segmentation and Landmark Detection Algorithms: VISCERAL Anatomy Benchmarks, IEEE Transactions on Medical Imaging, 2015.

[20] Henning Müller, Jayashree Kalpathy-Cramer, Allan Hanbury, Keyvan Farahani, Rinat Sergeev, Jin H. Paik, Arno Klein, Antonio Criminisi, Andrew Trister, Thea Norman, David Kennedy, Ganapati Srinivasa, Artem Mamonov, Nina Preuss, Report on the Cloud-Based Evaluation Approaches Workshop, SIGIR Forum, pages 35-41, 2016.

[21] S. K. M. Yi, M. Steyvers, M. D. Lee, M. J. Dry, The wisdom of the crowd in combinatorial problems. Cognitive Science 36(3):452–470, 2012.

[22] Henning Müller, Paul Clough, Thomas Deselaers, Barbara Caputo, ImageCLEF – Experimental evaluation of visual information retrieval, Springer, 2010.

[23] Dimitrios Markonis, René Donner, Markus Holzer, Thomas Schlegl, Roger Schaer, Georg Langs, Henning Müller, Khresmoi for radiology - Visual search in radiology archives and the open-access medical literature, IMAGING Management journal, 2013.

[24] John P. A. Ioannidis , Why Most Published Research Findings Are False. PLoS Med volume 2 number 8, 2005.

[25] Alex Krizhevsky, Ilya Sutskever, Geoffrey E. Hinton, Imagenet classification with deep convolutional neural networks. Advances in neural information processing systems. 2012.