Assignment 1: The Warm-Up

The **deadline** is May 3rd at 14:00 Saarbrücken standard-time. You are free to hand in earlier. You will have to choose **one** topic from the list below, read the articles, and hand in a **report** that critically discusses this material and answers the assignment questions. Reports should summarise the key aspects, but more importantly, should include original and critical thought that show you have acquired a meta level understanding of the topic – plain summaries will *not* suffice. All sources you use should be appropriately referenced, any text you quote should be clearly identified as such. The expected length of a report is between 3 to 5 pages, but there is no limit.

For the topic of your assignment, choose one of the following:

1. *The Scientific Discourse*

   Read [1], where the authors introduce a novel measure of correlation (or *dependence*) to detect associations in large data sets. The paper presents the findings in a very confident tone, but other researchers didn't entirely agree with [1]. Read [2] and [3] to see some of the rebuttals. The authors of [1] responded to [3] in [4] with the authors of [3] responding to [4] in [5].

   Who is right here? Is [1] presenting false claims and over-selling the results? Are the authors of [3] purposefully mis-interpreting [1] and ignoring their claims? Are the authors of [2] presenting sensible criticism, or should the note be completely ignored? What is your opinion, is the concept of *equitability* a useful one, and is *MIC* a useful measure? Can data mining benefit from these? Consider also the latest pre-print [6] by the authors of [1] in your answer.

   What does this exchange of letters and notes tell about the process of doing science? Has the general public been mis-led by some of these publications? Was the *Science* magazine wrong at publishing [1]? Should they retract it? Was it acceptable for *PNAS* to publish [3]? What is the value of pre-print servers such as arXiv (where [2] and [6] are published)?

   Your report should answer to both the technical questions and the above questions about the process.

2. *Deep Learning: The Best Thing Since Sliced Bread or Just Another Bottle of Snake Oil?*

   To answer to this assignment, you must have a good understanding of machine learning techniques in general, and deep learning techniques in special.

   The most talked-about data analysis topic of the last year or two has definitely been deep learning. Many researchers have claimed that they have obtained impressive results with deep learning: they can classify images [7], write LaTeX articles that almost compile [8], dream up images [9], and even win against the best humans on GO [10].

   Explain what is deep learning and how did the cited applications use various deep learning techniques. Do they all use one unified "deep learning algorithm"? If not, explain the differences and similarities between the approaches. Are they all really only about deep learning? For example, in your opinion, is AlphaGo primarily deep learning or reinforcement learning method? Did it do all feature selection automatically, or were there hand-crafted rules?

   And have all these applications really been resounding success stories? Read also [11] and [12]. How do they affect on your opinion on deep learning? Does deep learning have anything to do with data mining and knowledge discovery? Is deep learning the best thing since sliced bread, just another bottle of snake oil, ill-defined term under which people place all kinds of research, or what?

3. *I Want That One*

Data Mining aims at finding interesting novel knowledge from data. Yet, what is interesting? What is interestingness? De Bie [13,14] argues interestingness is inherently subjective, and that we should try to model the state of mind of the analyst, and proposes to do so using information theoretic principles. Discuss. Example questions you could address, but are not limited to include: What are the desired properties of the functions $InformationContent$ and $DescriptionalComplexity$? Is it not a play of words to hide *subjective* interestingness in two objective scores? Does it make sense to make the one subjective, and the other not? Can this be fixed? If so, how? How can we evaluate whether these functions behave well? That is, whether they correspond to what users find complex, resp. simple?

4. *Sounds... familiar...*

   Cilibrasi and Vitanyi define the ultimate clustering using Kolmogorov complexity [15] and show it can be approximated using off-the-shelf compression algorithms: you only need to use a data compression algorithm that suits your data. Is the framework they propose as powerful as they claim? Does this make sense at all? Why and when (not)? Are there any hidden assumptions? Can you think of examples where this framework would not be so easy to apply, or not work at all?

   Read [16]. What is the novelty of this paper w.r.t [15]? Are there key differences in the scores? What are the implications? How would you rate the novelty of this paper, what would you say is the key contribution?

   (Bonus) In a follow-up paper [17] Cilibrasi and Vitanyi propose to use *Google* as a `compression algorithm'. Does this make sense? Argue why (not).

Return the assignment by email to tada-staff@mpi-inf.mpg.de by **3 May, 1400 hours**. The subject of the email must start with [TADA]. The assignment must be returned as a PDF and it must contain your name, matriculation number, and e-mail address together with the exact topic of the assignment.

Grading will take into account both Hardness of questions, as well as whether you answer the Bonus questions.

References

You will need a username and password to access the papers outside the MPI network. Contact the lecturer if you don't know the username or password.

[1] Reshef, D.N., Reshef, Y.A., Finucane, H.K., Grossman, S.R., McVean, G., Turnbaugh, P.J., Lander, E.S., Mitzenmacher, M. & Sabeti, P.C. Detecting Novel Associations in Large Data Sets (../papers/a1/reshef11detecting.pdf). Science, 334(6062):1518-1524, 2011.

[2] Simon, N. & Tibshirani, R. Comment on Detecting Novel Associations in Large Data Sets by Reshef et al, Science Dec 16, 2011 (../papers/a1/simon11comment.pdf). arXiv, 1401(7645), 2011.

[3] Kinney, J.B. & Atwal, G.S. Equitability, mutual information, and the maximal information coefficient (../papers/a1/kinney14equitability.pdf). Proc. Natl. Acad. Sci. USA, 111(9):3354-3359, 2014.

[4] Reshef, D.N., Reshef, Y.A., Mitzenmacher, M. & Sabeti, P.C. Cleaning up the record on the maximal information coefficient and equitability (../papers/a1/reshef14cleaning.pdf). Proc. Natl. Acad. Sci. USA, 111(33):E3362-E3363, 2014.

[5] Kinney, J.B. & Atwal, G.S. Reply to Reshef et al.: Falsifiability or bust (../papers/a1/kinney14reply.pdf). Proc. Natl. Acad. Sci. USA, 111(33):E3364-E3364, 2014.

[6] Reshef, Y.A., Reshef, D.N., Sabeti, P.C. & Mitzenmacher, M.M. Equitability, interval estimation, and statistical power (../papers/a1/reshef15equitability.pdf). arXiv, 2015.

[7] Krizhevsky, A., Sutskever, I. & Hinton, G.E. ImageNet Classification with Deep Convolutional Neural Networks (../papers/a1/krizhevsky12imagenet.pdf). In NIPS '12, pages 1097-1105, 2012.

[8] Karpathy, A. The Unreasonable Effectiveness of Recurrent Neural Networks (http://karpathy.github.io/2015/05/21/rnn-effectiveness/). , 2015.

[9] Titcomb, J. Google unleashes machine dreaming software on the public, nightmarish images flood the internet (http://www.telegraph.co.uk/technology/google/11712495/Google-unleashes-machine-dreaming-software-on-the-public-nightmarish-images-flood-the-internet.html). , 2015.

[10] Silver, D., Huang, A., Maddison, C.J., Guez, A., Sifre, L., van den Driessche, G., Schrittwieser, J., Antonoglou, I., Panneershelvam, V., Lanctot, M., Dieleman, S., Grewe, D., Nham, J., Kalchbrenner, N., Sutskever, I., Lillicrap, T., Leach, M., Kavukcuoglu, K., Graepel, T. & Hassabis, D. Mastering the game of Go with deep neural networks and tree search (../papers/a1/silver16nature.pdf). Nature, 529(7587):484-489, 2016.

[11] Khurshudov, A. Suddenly, a leopard print sofa appears (http://rocknrollnerd.github.io/ml/2015/05/27/leopard-sofa.html). , 2015.

[12] Curtis, S. Google Photos labels black people as 'gorillas' (http://www.telegraph.co.uk/technology/google/11710136/Google-Photos-assigns-gorilla-tag-to-photos-of-black-people.html). , 2015.

[13] De Bie, T. An information theoretic framework for data mining (../papers/a1/tijl11kdd.pdf). In Proceedings of the 17th ACM International Conference on Knowledge Discovery and Data Mining (SIGKDD), San Diego, CA, pages 564-572, ACM, 2011.

[14] De Bie, T. Subjective Interestingness in Exploratory Data Mining (../papers/a1/tijl13ida.pdf). Springer, 2013.

[15] Cilibrasi, R. & Vitányi, P. Clustering by Compression (../papers/a1/cili05ncd.pdf). IEEE Transactions on Information Technology, 51(4):1523-1545, 2005.

[16] Campana, B.J. & Keogh, E. A Compression Based Distance Measure for Texture (../papers/a1/ck10mpeg.pdf). In Proceedings of the 10th SIAM International Conference on Data Mining (SDM), Columbus, OH, 2010.

[17] Cilibrasi, R. & Vitányi, P. The Google Similarity Distance (../papers/a1/cili07gsd.pdf). IEEE Transactions on Knowledge and Data Engineering, 19(3):370-383

(http://mmci.uni-saarland.de)