LUT University

MFG 4.0

Notes from BigDat2019 Winter School, Cambridge Jan 7-11th 2019

Jyrki Savolainen post-doc researcher at LUT

The development of data science and its methodologies are in the center of future's lean, automation-driven, manufacturing paradigm referred to as ***"Manufacturing 4.0"***. Tangible, production facility investments usually have a planned lifetime of at least 20 years giving you time to adapt with the technology whereas the life of data science technologies is usually a lot less.

For example, on-premises database solutions are still in a steady shift towards cloud computing and production critical is increasingly sold as a service instead of "fixed" licenses. This change is ongoing whether your ICT-department personnel (with their degrees dated in the 1990s) like it or not.

BigDat2019 Winter School was held at **University of Cambridge** Jan 7-11th

Richard **Bonneau** from New York University lecturing on: *Large Scale Machine Learning Methods for Integrating Protein Sequence and Structure to Predict Gene Function.*

Personally, I find it hard to keep myself updated on the latest developments of data science. To ease my life a bit I attended BigDat2019 winter school, where top researchers from different branches of data science shared their thoughts in a five-day lecture-marathon[1]. In my mind, the vast (freely available) online material does not resolve the problem of staying up-to-date. Simply, you just don't have time to search for "the good stuff" (your data pre-processing phase) and studying it thoroughly (data analysis). Even if you would manage to do these aforementioned steps, there's little if any guidance provided on whether you should predictively start to change your current ways of doing things (your data insights). Therefore, in this post I want to share my picks of the BigDat2019 winter school program and synthetize them into more general insights that hopefully are of interest for both the experts and nonprofessionals.

LUT University

MFG 4.0

More about MFG 4.0 Project

strategic RESEARCH

Turun yliopisto University of Turku

JYVÄSKYLÄN YLIOPISTO UNIVERSITY OF JYVÄSKYLÄ

HELSINGIN YLIOPISTO HELSINGFORS UNIVERSITET UNIVERSITY OF HELSINKI

SUOMEN AKATEMIA FINLANDS AKADEMI ACADEMY OF FINLAND

Notes from BigDat2019 Winter School, Cambridge Jan 7-11th 2019

Jyrki Savolainen post-doc researcher at LUT

Kenji Takeda, Director, Health and AI Partnerships, Microsoft Research:

A keynote speech was given by Microsoft's **Kenji Takeda**. Taking out the marketing part of it, he highlighted the complexity of real world versus the general perception of data science where sample, highly refined, datasets are downloaded from online coding competition platforms – such as *Kaggle* or *Codalab* – and then analyzed with ready-made pieces of code. On the contrary, the major part of the data scientist work still lies in the problem formulation and/or data preprocessing.

According to **Andrey Ustyuzhanin**, the power of data science competitions is the idea that a large problem can be divided into small pieces in a way that solving them does not require domain knowledge of the industry concerned.

For example, you were not required to have a physicist degree when taking part in the *Higgs Boson Machine Learning Challenge*[2]. So, do not sweat if you don't have the top-notch skills on algorithm optimization and testing (yet?), the effective use of crowdsourcing may leave the dirty work of algorithm testing and optimization for the masses rather than make you unemployed.

Andrey Ustyuzhanin from National Research University Higher School of Economics: *Challenge-driven Data Science: Cracking Domain Problems by Crowd Intelligence*

More about MFG 4.0 Project

Notes from BigDat2019 Winter School, Cambridge Jan 7-11th 2019

Jyrki Savolainen post-doc researcher at LUT

Data mining is a commonly used term, which in everyday language, refers to extracting meaningful insights from datasets (using statistical methods). Well, here's a new word: *process mining.* According to **Will van der Aalst**, process mining focuses in the analysis of business processes with the methods of data mining. If you want, you can regard it as a buzzword as this is what data analysts do already. Anyhow, in *process mining*, we are interested of chained discrete events, for example, person buying a concert ticket from an online store (event *A*) that is followed by receiving payment by invoice (*B*) or credit card (*C*) and ending in a successful delivery (*D*) or some other (unwanted) end-point (*E, F,…*). The recorded processes are therefore in form of combinations *"ABCD", "ACDF", "ABF"*, etc. Interestingly, this type of analysis usually reveals something that should

The point here, highlighted by van der Aalst, is that before getting your *"super-boosted-random-rainforest"* -algorithm running, check out what is really happening from the event data. There is a possibility that first having a clear view at your underlying business process (with deviations) may be actually the most valuable insight, which obsoletes any further analytics. To my surprise, the software available is simple to use and I strongly recommend to give one of them a test drive using provided tutorials[3]. Then you can make an informed decision whether it is any good for your analysis purposes.

Lastly, one technical note, as a wake-up call for the company representatives. *Kaggle, Microsoft Azure, Amazon Web Services* (among others) provide free *Jupyter Notebook* (or equivalent) interfaces to use *R* and *Python* for programming online. This means that instead of building your on-premises computing capacity and maintaining the software, you can directly access the latest development tools online running on state-of-the-art machines[4].

Wil van der Aalst from RWTH Aachen University
Process Mining: Data Science in Action

This is a huge benefit for those interested doing some testing on data analytics, but not yet willing to invest in the long-term solutions. I think most of the companies are still in this wait-and-see -phase when it comes to implementing some serious data analytics – especially in the SME-sector. For these people, I suggest going online and start tweaking with your data!

[1] For full program see: http://bigdat2019.irdta.eu/ (link tested on 12th Jan 2019)

[2]https://www.kaggle.com/c/higgs-boson

[3] ProM: http://www.promtools.org/

[4] For example https://www.kaggle.com/kernels (needs registration)

More about MFG 4.0 Project