A plan for future research

A headline for my plan for future research is: *further development of statistically based data science/big data methodology for application to geodata*. This should be carried out with a view to developing *generic* methodologies such that they may potentially be used across applications areas and not to geodata alone. A data scientist should think of data as a new raw material which should be applied to generate value within the application domain.

The accelerating development in applications of unmanned aerial vehicles, UAVs, also known as drones represents an opportunity for the (Danish and European) geoinformation industry, a field in which Denmark is traditionally strong. Also, an increasing number of satellites such as the European Sentinel series will deliver longer and longer global time series. We have

- global satellite optical image data coverage since the early 1970s,
- global satellite based sea surface height data since 1992,
- (global satellite based magnetic field data since 1999),
- global satellite based gravity data since 2002,
- global satellite polarimetric radar image data coverage emerging,
- Global Navigation Satellite Systems (GNSS) such as GPS and soon Galileo etc.

These developments leading to the advent of routinely collected (local and global) multi-source data make the appointment of a professor in data science and complexity within the geodata area very timely. Also, the potentially generic nature of the analysis methods needed makes the affiliation with DTU Compute highly relevant.

The above trends lead to an increased need for physically/mathematically/statistically based data science/big data methodologies, for example for mapping purposes, for the study of spatio-temporal dynamics including change detection, and for the derivation of information on important (for example climate) parameters from the data.

Said developments and trends create business opportunities and should be used for Denmark to retain and further strengthen its position as a leader in the global geoinformation industry, both as far as established companies and new businesses are concerned. Here DTU should facilitate the provision of BScs, MScs and PhDs within the area.

Danish obligations when it comes to monitoring for security, defense and climate purposes in Greenland and the Arctic as such are facilitated by some of the above data types and by data science methodology. This has just recently become even more relevant with the Danish presentation of a claim to the United Nations to an 895,000 km^2 area along the Lomonosov Ridge, an area covering the North Pole and also claimed by Russia.

Some of the above data types can be used to study global phenomena for example sea level rise and the El Niño phenomenon. El Niño is a very large-scale warm ocean event in the equatorial Pacific which can have huge socio-economic impacts globally (caused by for example drought in normally wet regions or torrential rain in normally dry regions). El Niño can affect global commodity prices and the macro-economy of different countries. Also, it may cause loss of lives: in the 1997-1998 El Niño event there were 21,000 estimated casualties (and more than US$ 22 billion in damage) world-wide.

The data mentioned are instrumental also in connection with natural disasters for example mud slides and earthquakes. Pre-event data are useful for establishing an inventory, and post-event data when compared with pre-event data can be used to assess damages and for planning of relief actions.

As a final geo-related example the use of airborne laser height measurements also generates enormous amounts of data. These data can be used to establish updated terrain height models which can be used to fight flooding caused by heavy rain or rising sea level. Also, better short-term weather forecasts may be useful in fighting flooding caused by heavy rain and other (anthropogenic or geogenic) catastrophes such as avalanches caused by among other things heavy precipitation.

Natural disasters, extreme weather, as well as sea level rise and El Niño may result in loss of life, damage of property and infra-structure, and consequently enormous economic costs for society locally and globally. Therefore their prediction, the dampening of their consequences, or perhaps ideally even their prevention are of paramount societal importance. Such challenges create business and job opportunities. Some of these challenges can be met by means of data science methodology.

Considerations as the above combined with the often used 3V-big-data-model (**v**olume, **v**elocity, **v**ariability) which is sometimes extended with a fourth V (**v**eracity) and rarely but importantly a fifth V (**v**alue), lead to a need for (and these bullets constitute my research vision)

- In general, further physically, mathematically and statistically based method development.
- Methodology development with a view to inter-disciplinary use.
- Computer implementation of methods with a view to data science/big data aspects, i.e., the handling of the enormous amounts of data collected routinely (in the geodata domain and in many other domains). These methods include parallel programming in clusters of CPUs using for example MPI, MapReduce, Hadoop, Spark, Tez and/or GeoWave, or in (clusters of) GPUs using for example CUDA.
- Further research into spatio-temporal dynamics of time series of global and regional satellite data, both optical and radar, and in other types of data.
- Analysis at segment or patch level (as opposed to pixel or single sample level).
- Development of methods to handle data with very different genesis and therefore with different statistical distributions. This could be based on information theoretical concepts such as entropy and mutual information.
- Visualization of results from complex analysis methods and models by means of indigenously developed methods, and for example Google Earth, NASA World Wind and Microsoft Bing Maps.
- Integration of methodology from different data science sub-disciplines such as (exploratory) data analysis, (multivariate) statistics, signal processing, image processing, time series analysis, information theory, chemometrics, data mining, machine learning etc.
- Collaboration between data scientists and subject-matter experts such as (geo)physicists, geologists, meteorologists, geographers, physicians, (bio)chemists, biologists etc.
- Spin-off and business development.

An important aspect here is the *generic nature* of the methodologies developed. This means that many methods although developed within the geodata realm, can inspire, or with little or no modification be applied directly to other subject-matter domains including life science, medicine, food quality, and industrial applications such as process and product control. This is a two-way street: methods developed in these latter domains may also inspire or potentially be carried more or less directly into the geodata world. This methodology exchange with cross-fertilization to and from other application domains, is extremely useful and should be exploited fully.

The above work can be performed in small informal projects (carried out as a part of the daily work) or in larger projects potentially funded by the European Commission's Horizon 2020 programme, by Copernicus, the European Earth Observation programme (formerly known as GMES, Global Monitoring for Environment and Security), or by Nordic or national research councils and foundations.

Some of the work should be carried out with a view to DTU's own COSINO project as reflected in the Danish report "Rummet kalder Jorden: Potentialet ved udvikling og anvendelse af nye satellitbaserede tjenester og produkter".

See also slides on this issue.