# Leiden Centre of Data Science

## Scientific Launching Symposium

Thursday 4 September 2014

Kamerlingh Onnes Gebouw
Room A051A
Steenschuur 25, Leiden

Universiteit
Leiden
The Netherlands

Academic Lectures
Scientific Launching Symposium LCDS
Leiden University - 4 September 2014

## Programme Academic Lectures LCDS

| | | |
|---|---|---|
| 14:00 | 14:05 hrs | Welcome by the Chair (Prof.dr. J.N. Kok) |
| 14:05 | 14:20 hrs | Data Theory is Statistical Science (or the other way around)<br>Prof.dr. A.W. van der Vaart |
| 14:20 | 14:35 hrs | Future of Biostatistics in Medical Research<br>mw. Prof.dr. J.J. Houwing-Duistermaat |
| 14:35 | 14:50 hrs | Astronomy and big data<br>Prof.dr. H.J.A. Rottgering |
| 14:50 | 15:00 hrs | Break |
| 15:00 | 15:15 hrs | Data science and the structure of cells and molecules<br>Prof.dr.ir. A.J. Koster |
| 15:15 | 15:30 hrs | Using 'big environmental data' to calculate the carbon-, water-, land- and material footprint of nations'.<br>Prof.dr. A. Tukker |
| 15:30 | 15:45 hrs | In silico Knowledge Discovery<br>From 'Data Integration' to Functional Interlinking<br>Prof.dr. B. Mons |
| 15:45 | 15:55 hrs | Break |
| 15:55 | 16:10 hrs | Genome biology beyond model organisms<br>Dr. C.V. Henkel |
| 16:10 | 16:25 hrs | Data Science and Network Theory: reconstructing the hidden linkages<br>Dr. D. Garlaschelli |
| 16:25 | 16:40 hrs | Volume versus complexity in simulation data<br>Prof.dr. S.F. Portegies Zwart |
| 16:40 | 16:45 hrs | Closing Address |
| 16:45 | 17:30 hrs | Drinks, Community Building |

## Data Theory is Statistical Science (or the other way around)

14:05 - 14:20 hrs **Prof.dr. A.W. van der Vaart**

Leiden University - Mathematical Institute

Statistical Science originated to answer simple questions, such as how to combine five measurements of the same quantity in a single number and quantify the remaining uncertainty. With a few more numbers things are different, but not completely so. We illustrate modern statistical reasoning with an example.

http://www.math.leidenuniv.nl/~avdvaart

## Future of Biostatistics in Medical Research

14:20 - 14:35 hrs **mw. Prof.dr. J.J. Houwing-Duistermaat**

Leiden University Medical Center - Department of Medical Statistics and Bioinformatics

In many studies, omics and Next Generation Sequencing data sets are available in addition to classical variables such as blood pressure. It is expected that integrated analysis of these various datasets will provide more insight into complex human traits. The joint analysis of multiple datasets requires a new statistical methodology. For an appropriate analysis of these datasets, statisticians need to understand the structure of the measurement errors of the novel datasets. Thus it necessitates communication with clinicians as well as with chemists. In the presentation I will show results and vistas of a number of multi dataset projects.

https://www.lumc.nl/houwing

## Astronomy and big data

14:35 - 14:50 hrs **Prof.dr. H.J.A. Rottgering**

Leiden University - Leiden Observatory

To understand the formation and evolution of galaxies, stars, and exoplanets and to address important questions related to the nature of dark matter and energy, huge astronomical surveys of the sky are currently been undertaken and/or planned. Collectively these surveys carried out either with ground or space-based facilities cover a significant fraction of the electromagnetic spectrum.

The computer challenges are often enormous. How to deal with the large data stream and complex analysis requirements need careful studies. In this talk we will give a brief overview of current and future surveys and associated big-data challenges. An emphasis will be on the projects that Leiden observatory is involved in.

http://home.strw.leidenuniv.nl/~rottgering/Site/Welcome.html

## Data science and the structure of cells and molecules

15:00 - 15:15 hrs **Prof.dr.ir. A.J. Koster**

Leiden University Medical Center- Department of
Molecular Cell Biology / Leiden University - NeCen

We will provide an overview of ongoing correlative light and electron microscopy approaches to image the structure of cells and molecules. We emphasize the recent technological developments and applications, in particularly those involved with digital data processing and visualization.

In the study of cellular processes two kinds of microscopy methods are combined: correlative light microscopy and electron microscopy. The idea is to allow a synergy between these two kinds of microscopy. Fluorescence light microscopy enables rapid searching for regions of interest in large fields of view and electron microscopy exhibits superior resolution over narrow fields of view. Fluorescence microscopy exploits the availability of a wide range of markers.

The integration of information (stemming from the two imaging modalities with large differences in spatial and temporal scale) provides novel insights into the fields of cells and structural biology (Faas et al, JCB, 2012). With fluorescence light microscopy, structures can be visualized. These structures have a resolution of several hundreds of nanometers, while electron microscopy can show molecular structures with only a few nanometer (Faas et al., JSB, 2013).

We will exemplify imaging of molecular structure by showing results of our work on the C1 complex, a molecular structure that plays a crucial role in the immune defense mechanism (Diebolder et al,, Science, 2014) against invading bacteria and viruses. We will also show 3D imaging results of cellular membranous structures that are induced during +RNA virus infection and are thought to serve as suitable microenvironments for viral RNA synthesis (Limpens et al., MBio, 2011).

https://electronmicroscopy.lumc.nl/index.html

## Using ‘big environmental data’ to calculate the carbon-, water-, land- and material footprint of nations’.

15:15 - 15:30 hrs **Prof.dr. A. Tukker**

Leiden University - Institute of Environmental Sciences

Many countries are very good at monitoring the resource extraction and emissions of production and consumption processes in their territory. It is however also very important to understand how these activities of production and consumption are connected. First, all production is ultimately driven by consumption. If we know the connections, we can understand how (1) changes in consumer behaviour, (2) changes in income patterns, and(3) changes in expenditure patterns will change the life cycle impacts of consumption. Second, we see that production chains have become global. Meat produced in Europe may be from husbandry fed with soy from Brazil. A car from a US manufacturer used in Argentina may contain electronics from China.

As a result, national resource extraction and emissions do not reflect the resource use and emissions related to final consumption in a country. There are countries that apparently meet internationally agreed emission reduction targets if we look at emissions within their boundaries, but at a closer look the life cycle emissions of their final consumption in fact has risen. Simply said: over time, production of emission-intensive products has moved from such countries to elsewhere.

We need hence an environmental accounting system that makes such relations visible between production, consumption, and related impacts, at a global level. So-called ‘Multi-regional Environmentally Extended Supply and Use/Input-Output Tables’ (MR EE SUT/IOT) are now widely seen as the most promising approach to create such an accounting system.

In the lecture we will give some intriguing environmental analytics, based on: economic supply and use tables of countries, production data per sector of countries. energy flow data from the International Energy Agency (IEA), Agricultural production data from the UN Food and Agricultural Organisation (FAO), emission data and emission factors, trade data.

http://www.cml.leiden.edu/organisation/staff/tukker.html

## In silico Knowledge Discovery
## From 'Data Integration' to Functional Interlinking

15:30 - 15:45 hrs **Prof.dr. B. Mons**

Leiden University Medical Center- Department of
Human Genetics

In the eScience era, the main activity of a data-driven Life Scientist is to combine relatively small datasets with the entire body of core legacy information. The main sources of information are literature, curated databases, and other datasets (GWAS, expression data, biobanks). As a result Knowledge Discovery (KD) becomes an excellent vehicle for pattern recognition in big data. An adequate tooling set for KD needs to address (1) the pattern recognition (Helicopter) phase named In Silico Knowledge Discovery, as well as (2) the specific discovery of patterns included in data created de novo, taken from a relatively small study. These two approaches constitute our major research line at LUMC and the Leiden Centre of Data Science. Other tooling needs that we investigate use confirmational reading by humans (the Excavation phase).
So far, attempts to create integrated tools to do In silico, de novo, and excavation have failed. We even believed that they are likely to fail forever. Actually, integrating massive datasets, usually conceived as Extract-transform- Load (ETL) processes of heterogeneous and dispersed datasets, is no longer an option. Currently, community driven consortia are beginning to show that modern semantic technology enables functional interlinking of massive datasets for everyone to use, as long as the de novo data talk the same language but still in relatively schema-free environments.

Examples will be given of how such functional interlinking efforts have already opened new avenues for exploration in big data sets such as LOVD, FANTOM5, HPA. All of them escape attention since Prople just uses an old fashioned literature search and shallow ETL. Thefuture lies in the data for life infrastructure. For instance, the infrastructure ELIXIR puts functional interlinking high on the European and global agenda's as it enables unprecedented data science.

http://www.biosemantics.org/index.php?page=barend-mons

## Genome biology beyond model organisms

15:55 - 16:10 hrs **Dr. C.V. Henkel**

Leiden University - Institute of Biology

Only a few years ago, large-scale genomics research was restricted to human biology and a few model organisms were of medical importance. Recently, DNA microarrays and 'first generation' DNA sequencing have been superseded by high-throughput ('next-generation') sequencing technologies. They are more flexible and less expensive. It has led to a democratization of genomics, bringing full genome sequencing and transcriptome analysis within reach of individual research groups working on non-model organisms.

The wider availability of genomics technology also compels many biologists to add computational methods to their repertoire of research methods, supplementing field observations and laboratory experiments. However, bioinformatics expertise remains a bottleneck. Currently, major challenges in bioinformatics include (1) the assembly of complete genomes from billions of tiny sequenced DNA fragments, (2) the extraction of meaningful patterns from large-scale gene expression profiling, and (3) the integration of results with data available in public repositories.

The combination of genomics and bioinformatics methods drives the integration of knowledge from many levels of biological science: from molecular biology, physiology and developmental biology, as well as from ecology and evolutionary biology.

In my talk, I will illustrate these developments using recent genomics-related results from the Institute of Biology, on such diverse research topics as venom evolution in snakes, migratory behaviour and sexual maturation in eels, plant genetic engineering, and models of human disease in zebrafish.

http://www.researchgate.net/profile/Christiaan_Henkel

## Data Science and Network Theory: reconstructing the hidden linkages

16:10 - 16:25 hrs **Dr. D. Garlaschelli**

Leiden University - Lorentz Institute for Theoretical Physics

Big Data are continuously produced. These data include privacy-protected information about individuals (e.g. social networks) and organisations (e.g. financial linkages). Publicly available data about social and financial networks represent only a fraction of the available data. Still, crucial decisions (by citizens, investors or banks) are typically made only on the basis of publicly available information. For instance, banks only disclose their total exposure towards the aggregate of all other banks, rather than their individual exposures towards each bank. However, how risk propagates across the banking system strongly depends on the detailed structure of the interbank network, which is typically known only to central banks. Is it possible to statistically reconstruct the hidden structure of a network, based only on partial (aggregate) information about its nodes? In this talk, I will give some answers to this question using a statistical physics point of view. I will present a recent enhanced reconstruction method and discuss its advantages and limitations. I will also comment on the possibility to detect early-warning signals of interbank collapse during financial crises.

http://www.lorentz.leidenuniv.nl/~garlaschelli/

## Volume versus complexity in simulation data

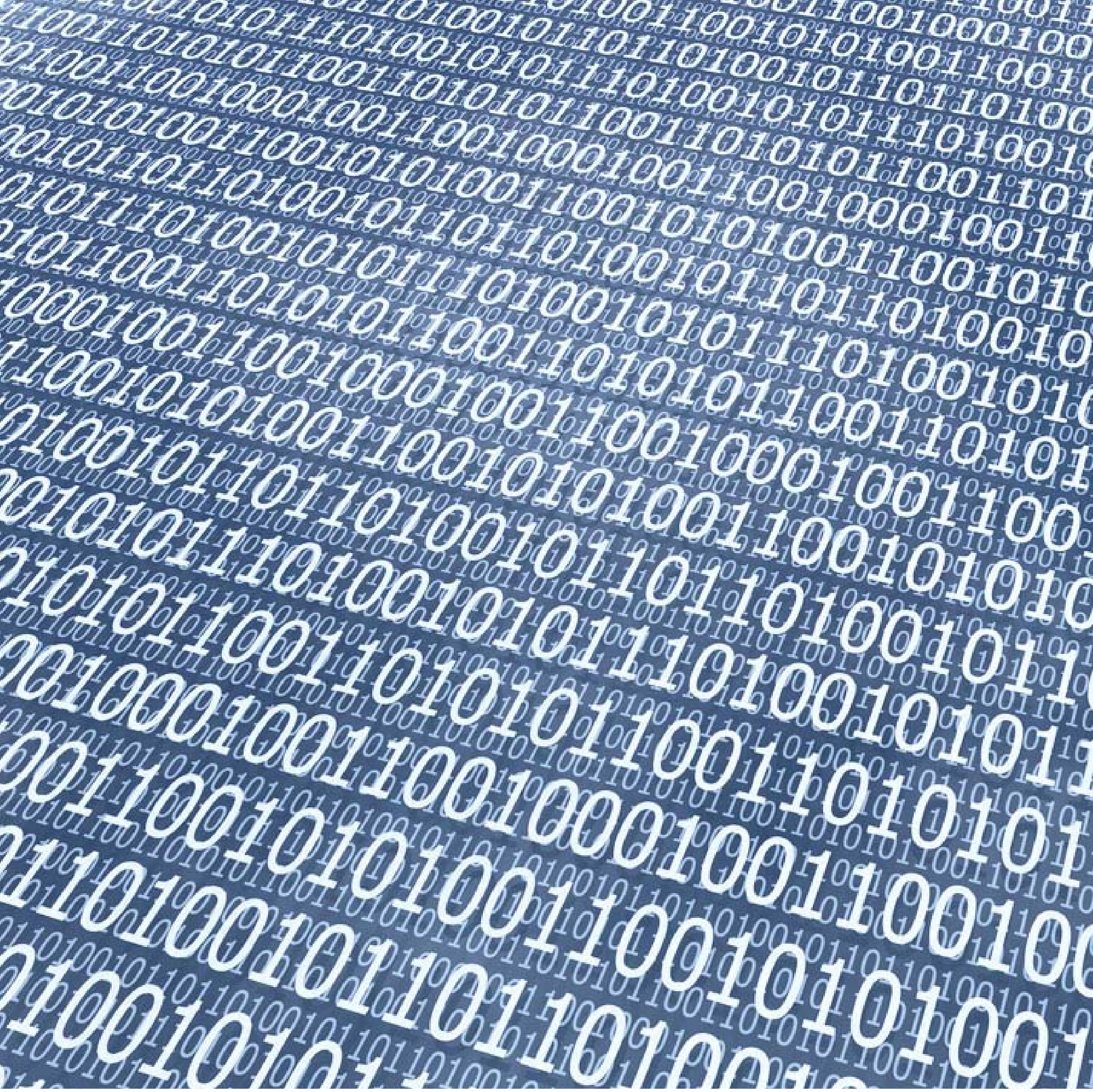
16:25 - 16:40 hrs **Prof.dr. S.F. Portegies Zwart**

Leiden University - Leiden Observatory

The complexity of simulations is in constant flux. For some astronomical simulations we have reached the point at which emergent behavior of the model has reached its natural equivalent. In principle we can even surpass nature using simulations, or simulate virtual environments in which the laws of physics are adapted. The volume and complexity of these data requires computer models to assist their interpretation. Will this vicious circle end or have we entered an eternal spiral?

http://home.strw.leidenuniv.nl/~spz/

## Notes



# Leiden Centre of Data Science

## Scientific Launching Symposium

Thursday September 4, 2014