Statistics and Data (Analytics): You cannot have one without the other

Jeanine Houwing-Duistermaat

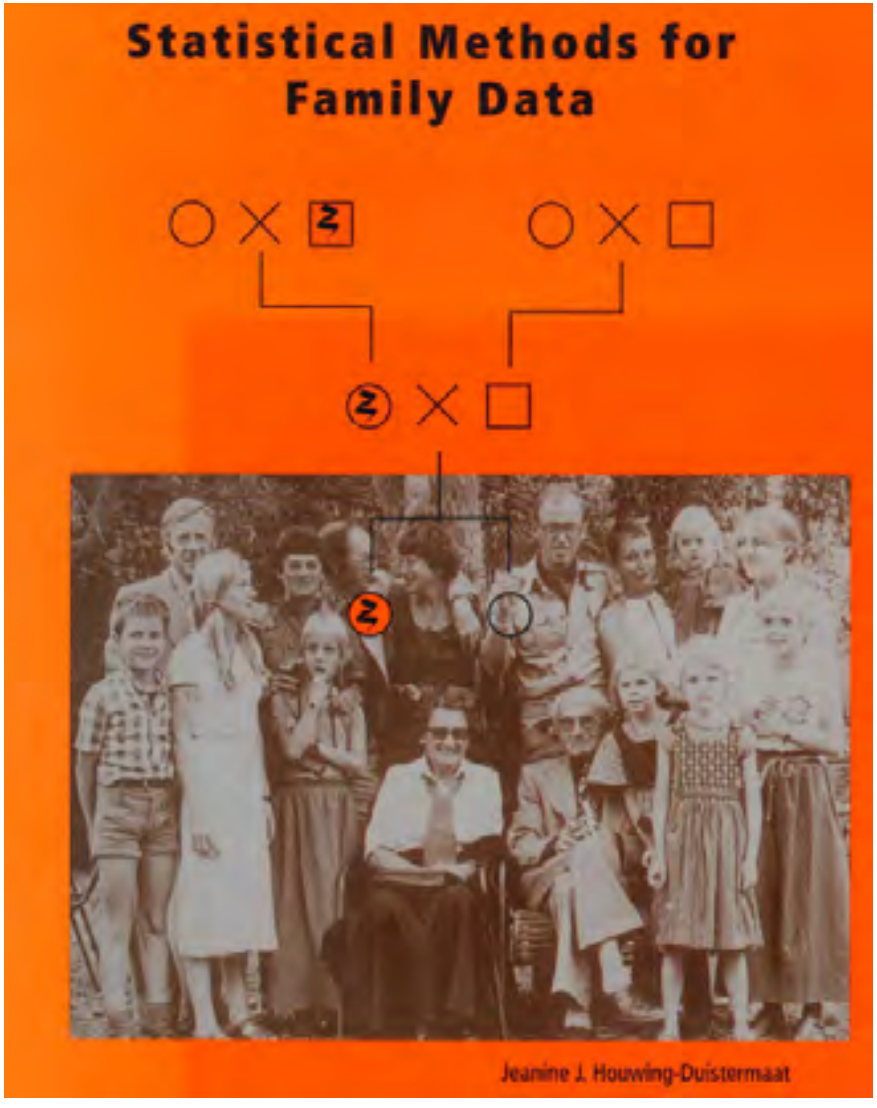
Chair in Data Analytics and Statistics
School of Mathematics

14 June 2016

UNIVERSITY OF LEEDS

Copenhagen, 1991-1992

PhD Thesis, 1992-1997



Collaboration with Maria Yazdanbakhsh

MODELLING THE CAUSE OF DEPENDENCY WITH APPLICATION TO FILARIA INFECTION

JEANINE J. HOUWING-DUISTERMAAT[1*], HANS C. VAN HOUWELINGEN[2] AND ANNEMARIE TERHELL[3]

Parasitology (2000), **120**:23-29 Cambridge University Press

Research Article

Clustering of *Brugia malayi* infection in a community in South-Sulawesi, Indonesia

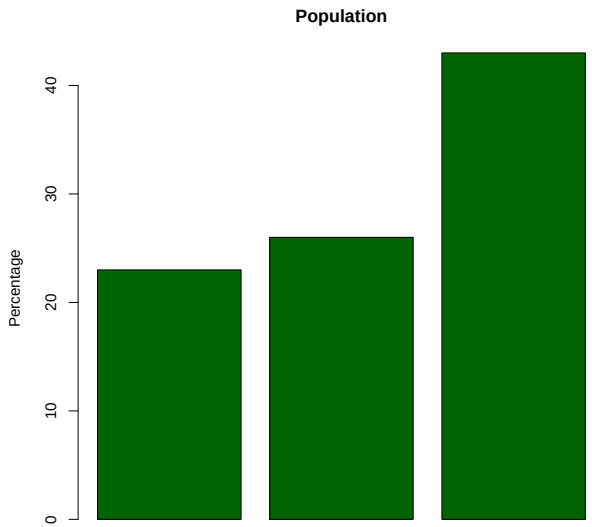
A. J. TERHELL [a1 c1], J. J. HOUWING-DUISTERMAAT [a2], Y. RUITERMAN [a1], M. HAARBRINK [a1], K. ABADI [a3] and M. YAZDANBAKHSH [a1]

JUMP!

2007 onwards

Families and twins with Rheumatoid Arthritis (RA)

- Collaboration Immunohematology and Blood Transfusion, Rheumatology and Statistics
- Funding for estimation of contribution of environmental and genetic factors to RA from family and twin studies (PI: Prof R de Vries)
- Statistical Challenges:
  - Model has to represent general population
  - Outcome dependent sampling
- Other challenges: Limited statistical knowledge and research goal which could not be accomplished

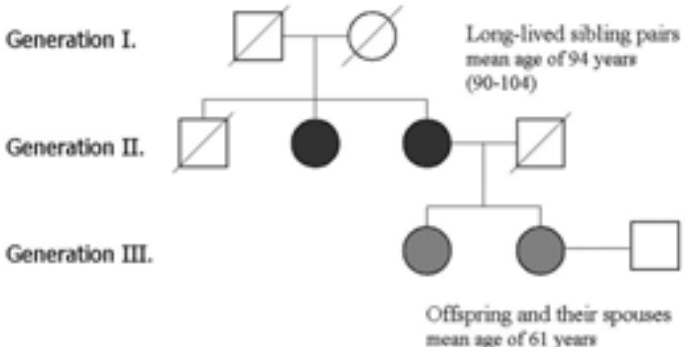


Telling figures indeed, but what do they mean to you, what do they mean to me, what do they mean to the average man in the street?
With me now is Professor Tiddles of Leeds University...
I think it's too early to tell.

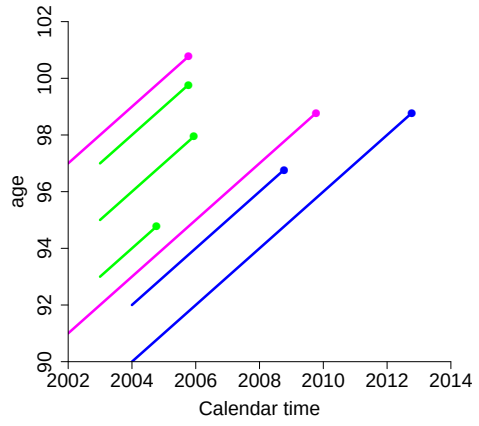
Monty Python

'Too early to tell' ... too early to say... it means the same thing. The word 'say' is the same as the word 'tell'. They're not spelt the same, but they mean the same. It's an identical situation, we have with 'ship' and 'boat' but not the same as we have with 'bow' and 'bough', they're spelt differently, mean different things but sound the same. But the real question remains. What is the solution, if any, to this problem? What can we do? Where do we stand? Where do we sit? Where do we come? What do we think? What do we do? What do they mean etc.

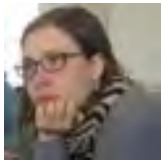
Fun: Modelling survival in long lived families



Leiden Longevity Study
PI: Prof P. Slagboom



- Statistical challenges:
  - Correlation among family members (frailty models)
  - Family members enter the study at different ages (delayed entry)
  - Within family missingness
- Idea: combine existing approaches with inverse probability weighting



Rodriguez Girondo et al 2016 Survival analysis with delayed entry in selected families with application to human longevity *Statistical Methods in Medical Research*

Big Data

- To summarize
  - Models for RA and Mortality in the elderly
  - Outcome dependent sampling
  - Missingness and Inverse Probability Weighting
- Nested multi-case family design?
- Weights from Electronic Health Records (EHR)?
  - Leiden: link databases of General Practitioners with community registration to obtain family structures
  - Include family information in EHR
- Big data, Statistics and Computer Science

www.universiteitleiden.nl/en/research-focus-areas/health-prevention-and-the-human-life-cycle

Fun: ImmunoSPIN and SugarSPIN

Scientific Program Indonesia - Netherlands

Prof Maria Yazdanbakhsh

ImmunoSPIN and SugarSPIN

- Colaboration between LUMC and University of Djakarta
- ImmunoSPIN (2007-2011): Parasitology, Immunology

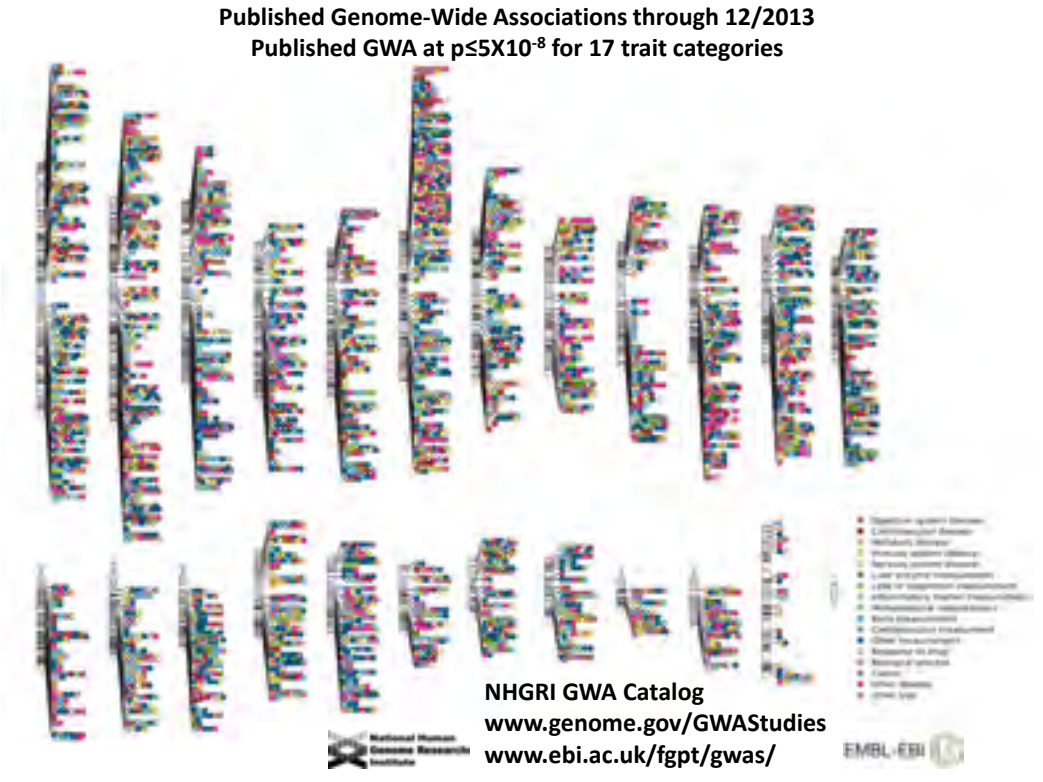
- SugarSPIN (2012-2017): Endocrinology, Immunology, Behavioral Science, Statistics
- Helminth infections and type 2 diabetes in Indonesia
- Statistical challenges: Joint analysis of mixed outcomes

Big data in Health

- **Electronic Health Records**
- **Omics:** Genomics, Epigenomics, Transcriptomics, Proteomics, Metabolomics, Microbiomics
- Clinical trials and cohort studies
- Patient registries
- Medical Imaging
- Data from devices
- Social Media, Occupational information, Environmental Monitoring etc

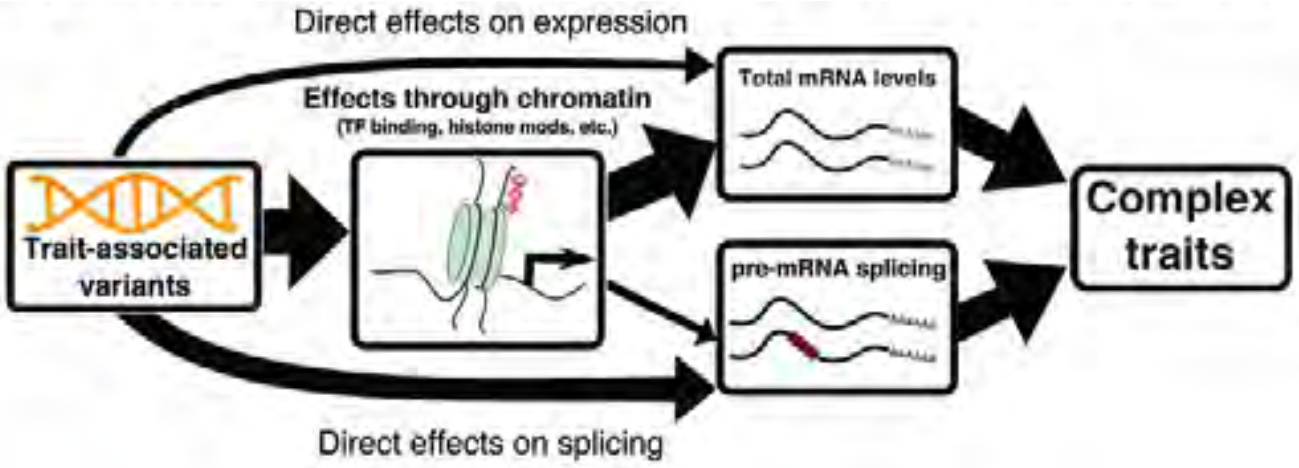
Statistical Genetics

- My first large dataset: Genome Wide Association studies



- Biological explanation?

Transcriptomics



Li et al 2016 Science RNA splicing is a primary link between genetic variation and disease

- Chromatin processes
  - Transcription starting sides, enhancer, silencer (3D)
  - Supercoiling (Dynamic)
- Statistical challenges
  - Integration of data and combining information
  - Measurement error
  - Modelling complexity of time and space (Sarah Harris, University of Leeds)

TopBreed

- Towards precision breeding using genomic prediction (2016-2019)
- Collaboration with Biometris & Animal Breeding, Wageningen University (Prof Fred van Eeuwijk, Prof Roel Veerkamp)
- To improve genomic prediction for phenotypes by using different types of genetic variation, information across generations and populations

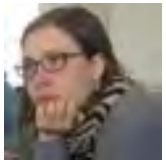
MIMOmics

Methods for integrated analysis of multiple omics datasets

- Genomics, Transcriptomics, Glycomics, Metabolomics
- 15 partners, around 50 scientists
- Expertise:
  - High Throughput Methods, Quantitative Methods, Epidemiology and Biology
  - Biostatistics, Machine Learning, Bio-informatics, Biophysics
  - Computing and Data Infrastructure
- Cartoon about MIMOmics can be found here:

  `http://www.jeanineHouwing.com/how-methods-for-data-analysis-improve-the-lives-of-john-and-suzy/`

Multi omics methods development in MIMOmics

- Integration of datasets and dimension reduction
  - Partial Least Squares 
  - Network approaches: Biostatistics & Biophysics 
- Prediction
  - Augmented value of omics datasets 
  - Prediction & Interpretation & Machine learning 
- Causal Inference together with Manchester
- Next step: multi omics in family studies

Big data and Precision Medicine

UNIVERSITY OF LEEDS

Methodological Framework for Family Data

- Study Design: Obtain information on structure and prevalence of disease within families from GP databases? (Owen Johnson, University of Leeds)
- Modelling of multiple omics data in families to obtain biological insight (extending MIMOmics)
- Family members share genetic and environmental factors which play a role in disease onset, hence family data should be used for risk prediction. (Jenny Barrett, University of Leeds)

Promise of big data in Health

- Precision medicine
- More and better research output
- Gain in efficiency and reduce costs

- Innovation

  past new question - new data

  now new question - can we use existing data?

  also data - which relevant questions can we answer?

Big data in Health

- **Electronic Health Records**
- **Omics:** Genomics, Epigenomics, Transcriptomics, Proteomics, Metabolomics, Microbiomics
- Clinical trials and cohort studies
- Patient registries
- Medical Imaging
- Data from devices
- Social Media, Occupational information, Environmental Monitoring etc

Methodological Contributions

- Study Design: Sampling informative parts from big data
- Better models for rare disease, common diseases and co-morbidity
- Statistics, Data Analytics, Computer Science
- Challenges?

Wiki definitions

Statistics

Statistics is the study of the **collection, analysis, interpretation, presentation, and organization of data**. In applying statistics to, e.g., a scientific, industrial, or social problem, it is conventional to begin with a **statistical population** or a statistical model process to be studied. Populations can be diverse topics such as "all people living in a country" or "every atom composing a crystal". Statistics deals with all aspects of data including the planning of data collection in terms of the design of surveys and experiments.

Analytics

Analytics is the **discovery, interpretation, and communication of meaningful patterns in data**. Especially valuable in areas rich with recorded information, analytics relies on the simultaneous application of statistics, computer programming and operations research to quantify performance. Analytics often favors data **visualization** to communicate insight.

American Statistical Association

What does statistics bring to Big Data and where are the opportunities

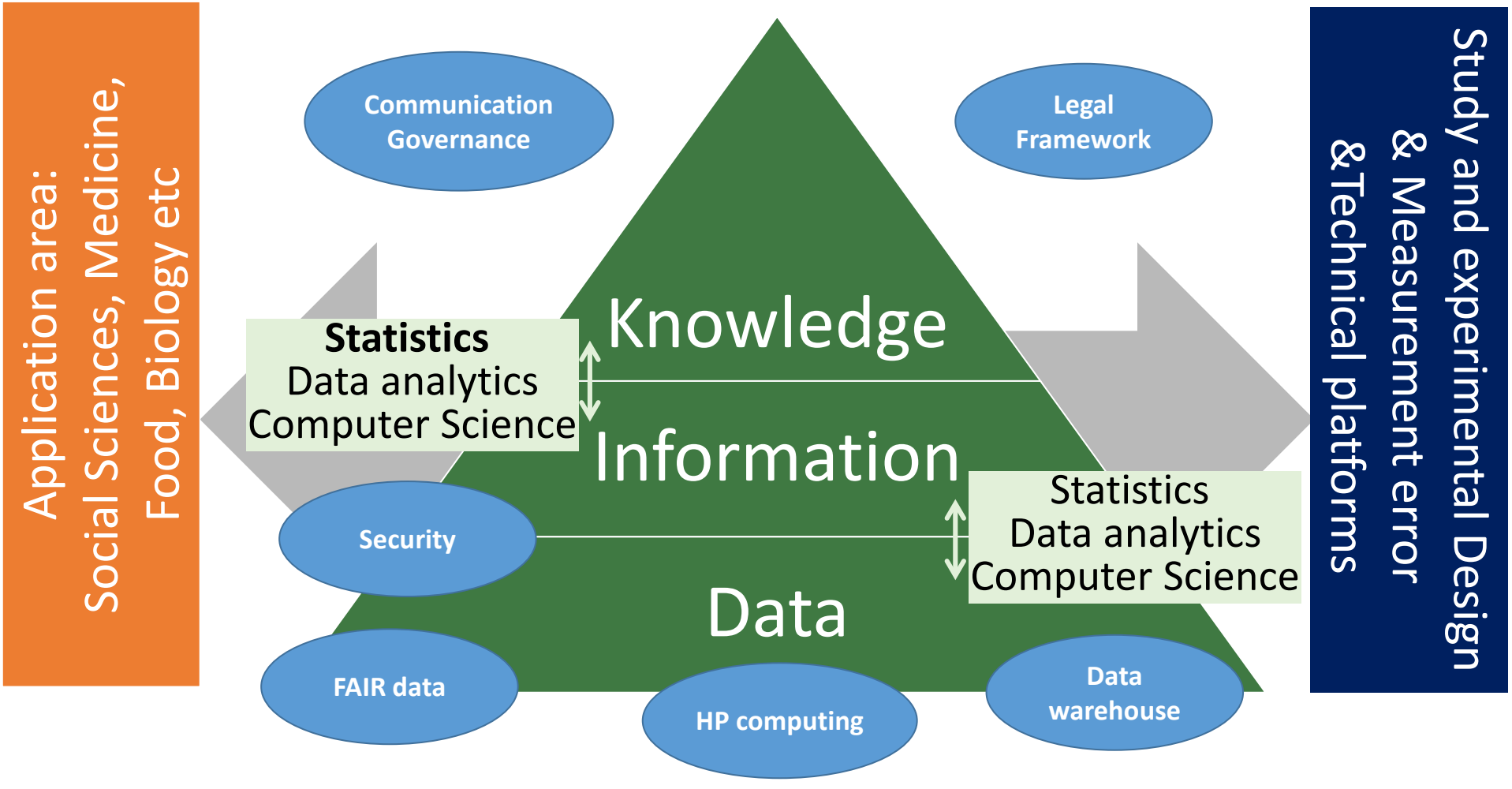
Statisticians are skillful at **assessing and correcting for bias**; **measuring uncertainty**; designing studies and **sampling strategies**; assessing the quality of data; enumerating limitations of studies; dealing with issues such as **missing data** and other sources of non-sampling error; developing **models for the analysis of complex data structures**; creating **methods for causal inference** and comparative effectiveness; eliminating redundant and uninformative variables; **method for combining data sources** and determining effective data visualization techniques;

and

Statisticians are used to collaborate with experts of **application areas**

www.amstat.org

Data Science

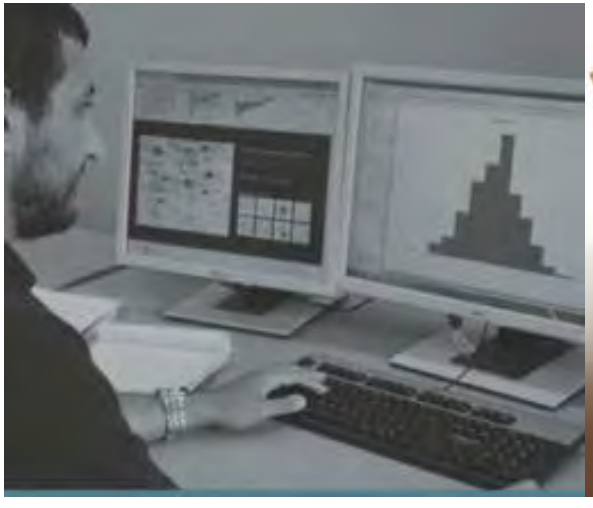


Why is Health Sector behind?

- Infrastructure
- Biology is difficult
- Integrate diverse datasets
- Preference for simple statistics
- Clinical Trials vs Observational Studies
- Multidisciplinary component is huge

Auffray C et al 2016 Making sense of big data in health research: Towards a European Union action plan. *Genome Medicine*

Multidisciplinary: expertise and skills



Multidisciplinary: Communication

Multidisciplinary: Commitment

Multidisciplinary: Innovation

Multidisciplinary: Leadership and vision

To conclude

- Methodological approaches for Big Data in Health
  - Study Design and Finding informative parts: Statistics, Computer Science, Data Analytics
  - Big data: Data Analytics, Statistics, Computer Science
  - Integration of diverse datasets
- Innovation
  - Data - which relevant questions can we answer?
- Multidisciplinary challenges
- Leadership, vision, dissemination

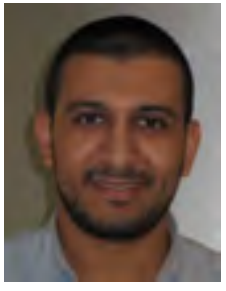
Thank you

Thank you

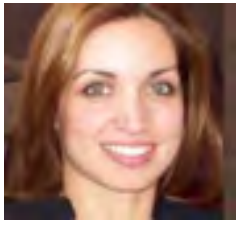
Statistical Genetics, LUMC





Georgios Bartzis,
Renaud Tissier,
Kate Xu,
Mar Rodríguez-Girondo,
Angga Fuady,
Said el Bouhaddani,
Ivonne Martin,
Hae Won Uh





Alexia Kakourou,
Bart Mertens,
Stefan Bohringer
Roula Tsonaka,
Szymon Kielbasa,
Ramin Monajami

Equality and Inclusion

Data & Methods & Multidisciplinary

Statistics