STRATOS Topic Group 9

Analysis of high-dimensional data: Opportunities, challenges and goals

Jörg Rahnenführer

Technische Universität Dortmund, Fakultät Statistik

Email: rahnenfuehrer@statistik.tu-dortmund.de

Seminar, LMU München

München, 05.12.2018

STRATOS

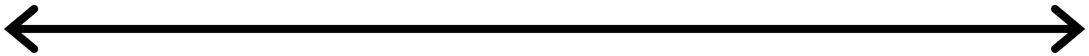
- An efficient way to help researchers *to keep up with recent methodological developments* is to develop guidance documents that are spread to the research community at large.

- The objective of STRATOS is to *provide accessible and accurate guidance* in the design and analysis of observational studies.

Analysis of high-dimensional data

- Situation: Many more variables than samples: p >> n
- Prediction models (regression, classification, survival):
  Inherent model selection problem

Bias/Variance – „Model fit“ vs. „Model complexity“



1 gene

50.000 genes

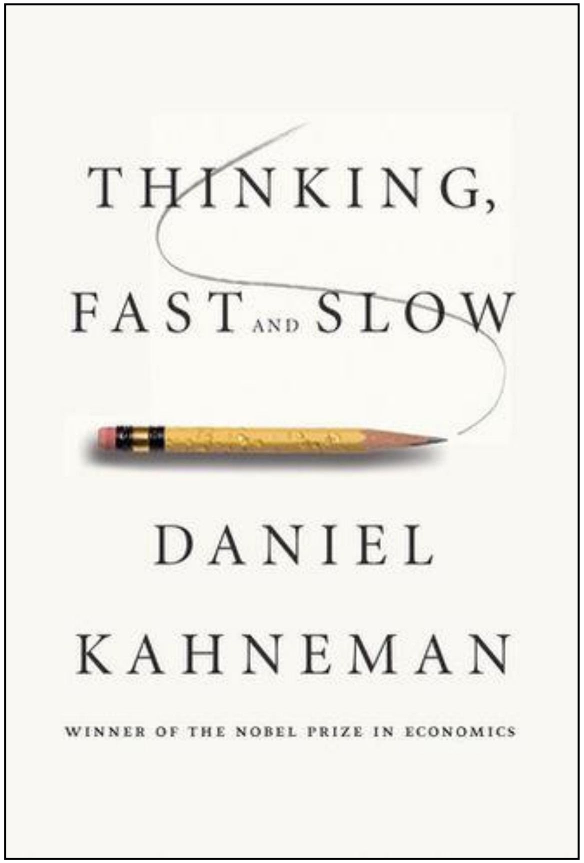
- Solutions for high-throughput data with variable selection
  - Filtering: Select “best” variables before modelling
  - Wrapping: Select variables “within” modelling algorithm

Analysis of high-dimensional data

- Joy of the analysis of HDD
  - Having fun with data – interesting and challenging
  - Great interdisciplinary research opportunities
- Reality check
  - In practice hard to always do what should be done
  - Much flexibility and tuning possible for the analysis of HDD data
- Pressure to publish – publication bias
  - Ioannidis, John P. A. (August 1, 2005). "Why Most Published Research Findings Are False". PLoS Medicine. **2** (8): e124
  - “Proteus phenomenon”: Occurrence of extreme contradictory results in the early studies performed on the same research question

Analysis of high-dimensional data

- Humans are definitely not good in avoiding pitfalls: Thinking – fast and slow (Daniel Kahnemann)
  - Confirmation bias: The tendency to search for, interpret, favor, and recall information in a way that confirms preexisting beliefs or hypotheses
  - Overfitting of numbers and patterns…



Stratos: Topic Groups

STRATOS
INITIATIVE

TG 1: Missing data

TG 2: Selection of variables and functional forms in multivariable analysis

TG 3: Initial data analysis

TG 4: Measurement error and misclassification

TG 5: Study design

TG 6: Evaluating diagnostic tests and prediction models

TG 7: Causal inference

TG 8: Survival analysis

TG 9: High-dimensional data

Motivation for TG 9

- Increasing use and availability of health-related metrics
  - Omics data (genomics, transcriptomics, proteomics, …)
  - Electronic health records
- Big data / high dimensionality
  - Big data typically refers to very large sample size n
  - High-dim: number of unknown parameters p is of much larger order than sample size n (p >> n)
- Problems
  - Heterogeneity (e.g., different sources, technologies)
  - Noise accumulation (accumulation of estimation errors)
  - Methods established for low-dim break down for high-dim!

Current goals for TG 9

- Overview paper
  - Statistical analysis of biomedical HDD:
    Main scenarios, common approaches and future directions
- Simulation paper
  - Guidance for planning, conducting and reporting simulation studies for comparing analytic approaches for biomedical data:

    General concepts with additional considerations for high-dimensional data
- Guidance for analysis processes
  - Examples for data analysis processes for specific types of HDD
  - Recommendations for best practices
  - R-Code with interpretations

TG 9: Members

- Federico Ambrogi (University of Milan, Italy)
- Axel Benner (DKFZ Heidelberg, Germany)
- Harald Binder (Freiburg University, Germany)
- Anne-Laure Boulesteix (LMU Munich, Germany)
- Tomasz Burzykowski (Hasselt University, Belgium)
- Riccardo De Bin (University Oslo, Norway)
- W. Evan Johnson (Boston University, USA)
- Lara Lusa (University of Ljubljana, Slovenia)
- **Lisa McShane (NCI, USA)**
- Stefan Michiels (University Paris-Sud, France)
- Eugenia Migliavacca (Nestle Institute of Health Sciences Lausanne, Switzerland)
- **Jörg Rahnenführer (TU Dortmund, Germany)**
- Sherri Rose (Harvard Medical School, USA)
- Willi Sauerbrei (Freiburg University, Germany)

STRATOS
INITIATIVE

TG 9: Subtopics

1. Data pre-processing
2. Data reduction
3. Exploratory data analysis
4. Multiple testing
5. Prediction modeling/algorithms
6. Comparative effectiveness and causal inference
7. Design considerations
8. Data simulation methods
9. Resources for publicly available high-dimensional data sets

Subtopic 1: Data Preprocessing

- Omics data: Removal of systematic biases
  - Intensity effect, batch effect, dye effect, block effect, …
- Challenges
  - Keeping up with new technologies that generate new data types
  - Methods specific to technology or data generating mechanism (NGS, single-cell transcriptomics, mass spectrometry)
- Typical Tasks
  - Normalization/calibration, identification of outliers/errors, transformations
- Newer approach
  - Build models where preprocessing is already part of the analysis process, not clearly separated

Subtopic 2: Data reduction

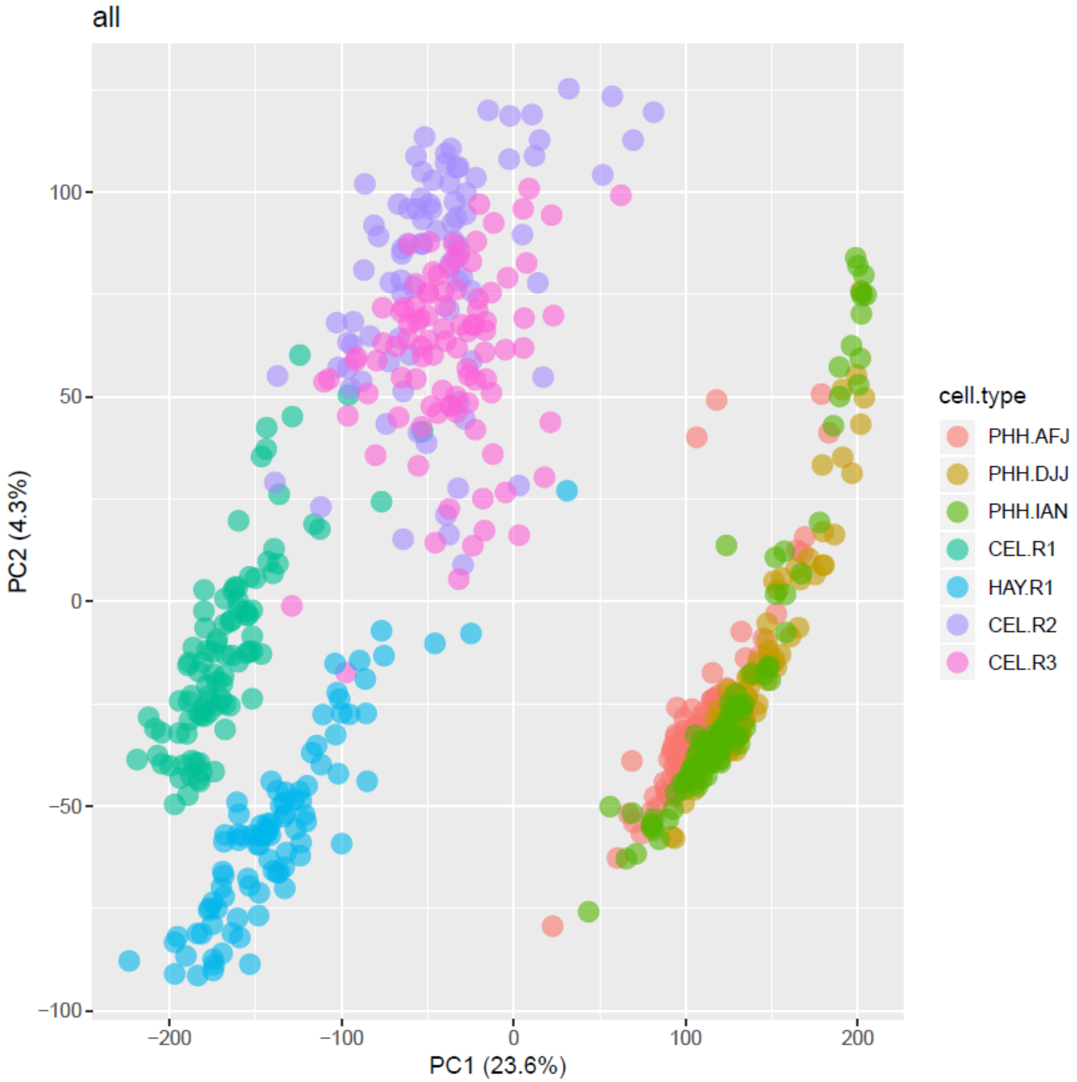
- Dimension reduction and variable selection
  - Central role for analyzing high dimensional data, in terms of statistical accuracy
- Goals
  - Building / finding prototypical samples
  - Building new variables, e.g. meta-genes, for use in subsequent statistical modeling or machine learning approaches
  - Cluster analysis, with subsequent aggregation of clusters
- Newer approaches
  - Projection to lower-dimensional space: t-SNE
  - Neural networks

Subtopic 2: Exploratory analysis

- Quality control
  - Identify potential problems and biases in the data, like batch effects, outliers, missing values etc.
  - Analysis of distributions of samples (across features)
- Grasp the structure of the data
  - Summary statistics – scores
  - Data visualization
    - Heatmaps
    - Projections into fewer dimensions: PCA, tSNE, …
  - Cluster analysis
    - Classical (k-means, …) and high-dim (subspace clus., DBSCAN)
    - Identify regions with relatively large data density
    - Biclustering

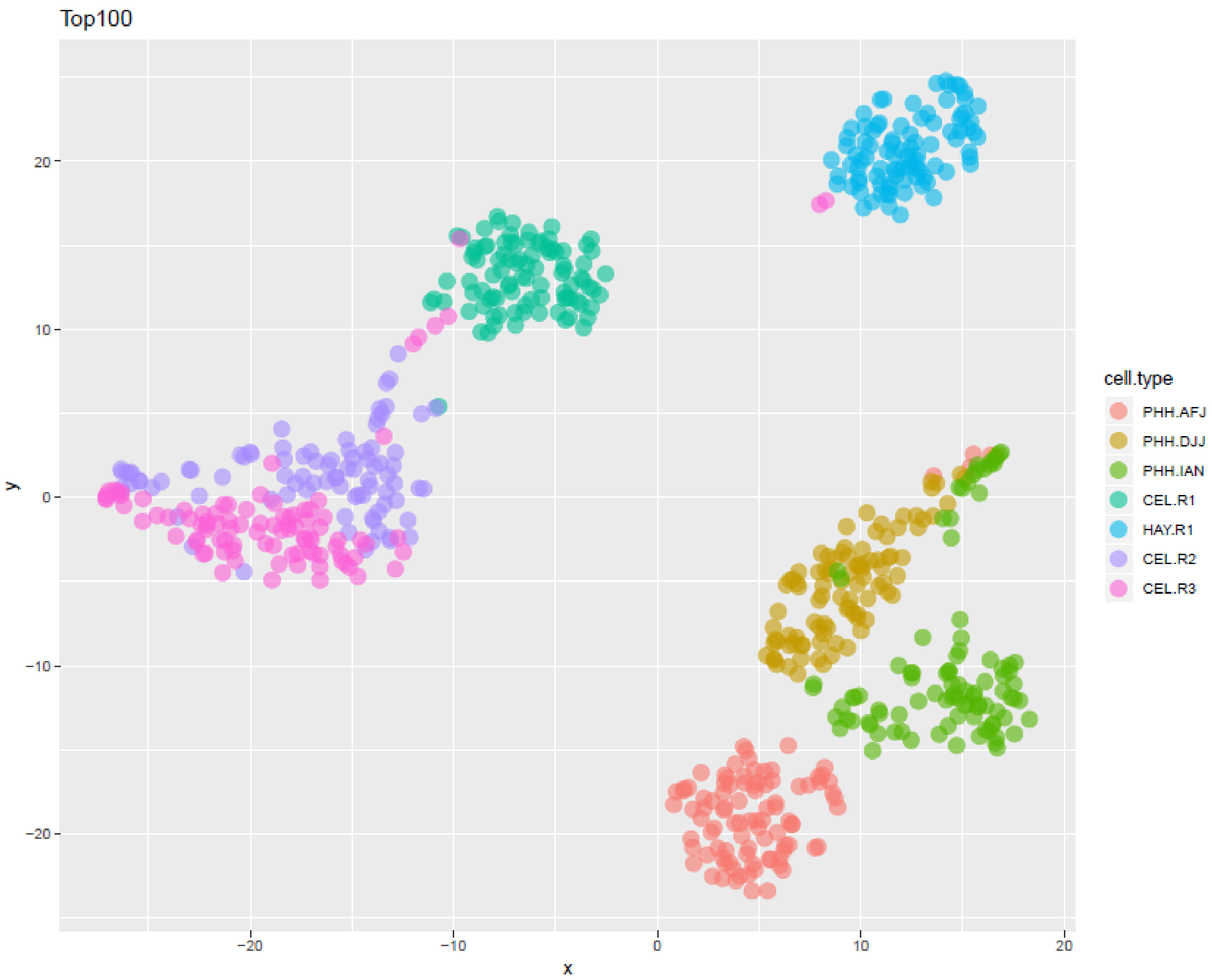
Subtopic 3: Exploratory analysis

- PCA plot for single-cell transcriptomics
- Color represents experiment
  - 96 values per experiment
- Left
  - 4 experiments with hepatocyte-like cells differentiated from human pluripotent stem cells
- Right
  - 3 experiments with cells from primary human hepatocytes



Subtopic 3: Exploratory analysis

- t-SNE plot for single-cell experiments
- Color represents experiment
  - 96 values per experiment
- Left
  - 2 replicates separated
- Right
  - 3 replicates clearly separated



Subtopic 3: Exploratory analysis

- DBSCAN
  - finds clusters of arbitrary shape, is robust to noise, and scales well to large databases (Ester, Kriegel, Sander, Xu, KDD 1996: A Density-Based Algorithm for Discovering Clusters in Large Spatial Databases with Noise)
- 2014 SIGKDD Test of Time Award
  - recognizes outstanding papers from past KDD Conferences beyond last decade with important impact on the data mining research community

http://www.kdd.org/News/view/2014-sigkdd-test-of-time-award

- Popular algorithm in computer science and data mining
  - but not much applied in statistics community, although successful/competitive in many applications
  - for example applied to clustering mass spectra (own research)

Subtopic 4: Multiple testing

- Statistical testing of thousands of hypotheses
  - Requires alternative procedures to control false discovery rates and to improve power of the tests
- Many different scenarios
  - Find variables with different distributions between pre-specified classes of subjects or with association with outcome
  - Enriched variables classes in a list of selected variables
- Statistical approaches
  - Control of false positives (e.g., FDR, empirical Bayes)
  - Global testing versus one-at-a-time testing
  - Enrichment tests (e.g., gene set enrichment analysis)

Subtopic 5: Prediction modelling/algorithms

- Differentiation between predictive accuracy and interpretation
- Prediction models
  - Binary/categorical (response to therapy)
  - Continuous (tumor size after therapy)
  - Survival (overall survival, disease free survival)
  - Interpretation of prediction model (parameters)
- Why standard methods break down
  - For n<<p cannot fit standard regression model
  - Redundancy in variables (huge correlation as problem for stable variable selection)

Subtopic 5: Prediction modelling/algorithms

- Model building
  - Penalized regression (ridge, lasso, elastic net, SCAD, MCP)
- Machine learning methods
  - Trees, support vector machines, multilayer neural networks
  - Random forests, boosting
  - Neural networks
- Evaluation of prediction models
  - Performance metrics (e.g., MSE, AUC, Brier score)
  - Risk of overfitting (consider stability, validation)
  - Tuning hyperparameters (nested CV)
  - Improper evaluation (e.g., resubstitution) drastically overestimates model performance (and is still extremely common)

Subtopic 8: Data Simulation Methods

- Issues specific to high-dimensional data
  - Underlying (biological) mechanism not well understood
  - Difficult to simulate realistic correlation structure and suitable multivariate distributions
- Approaches
  - Simulations based on assumed distributions (e.g. normal, Poisson, negative binomial)
  - Simulation using extracted parameters from pilot data
  - Simulation using real data (e.g., plasmode data)
- Plasmode approach
  - Plasmode (from plasm=form, and mode=measure) is a real (i.e., from actual biological specimens) data set for which some aspect of the truth is known (Mehta et al., Physiological Genomics, 2006)

Definition of deep learning

- “Deep learning is part of a broader family of machine learning methods based on learning representations of data”.

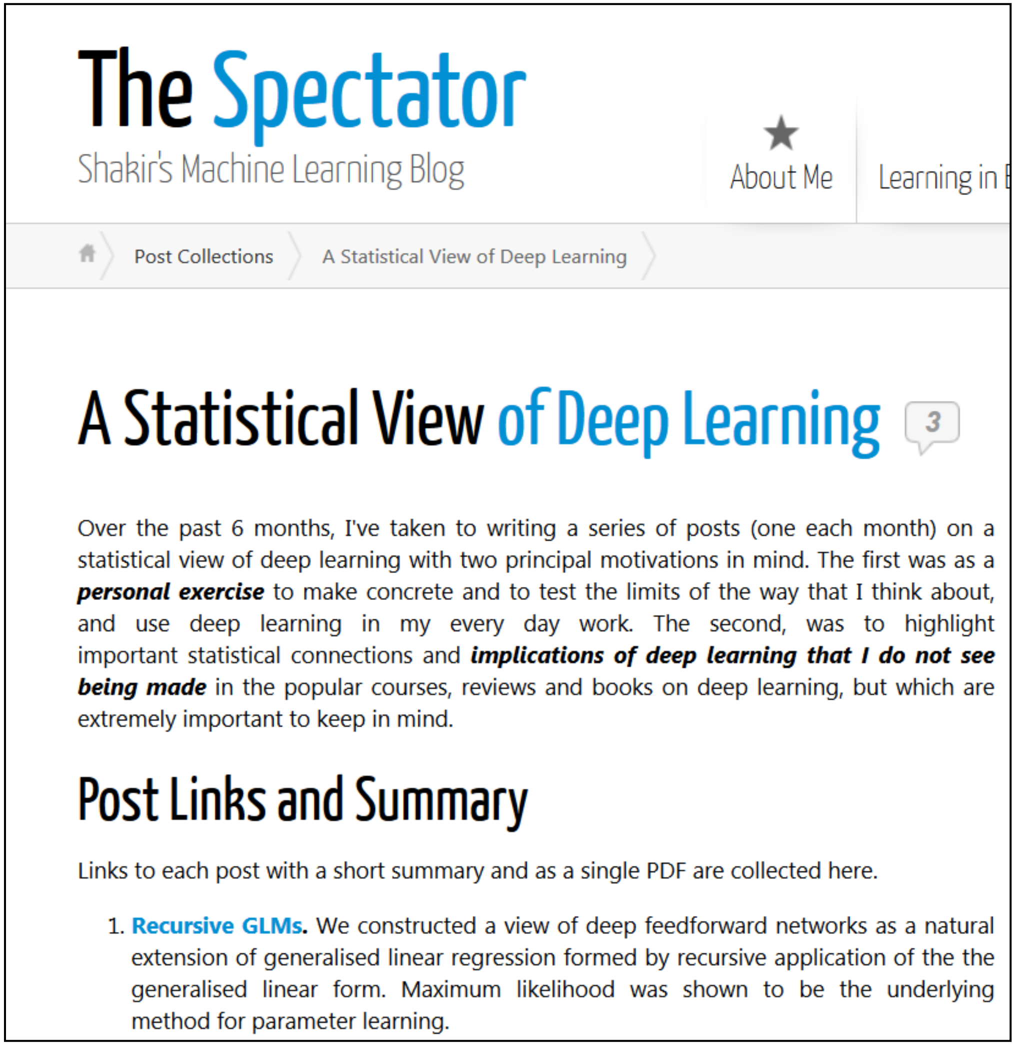
  Source: Wikipedia

- Idea
  - “An observation (e.g., an image) can be represented in many ways such as a vector of intensity values per pixel, or in a more abstract way as a set of edges, regions of particular shape, etc.”
  - “One of the promises of deep learning is replacing handcrafted features with efficient algorithms for unsupervised or semi-supervised feature learning and hierarchical feature extraction.”
- Types
  - “Deep neural networks, convolutional deep neural networks, deep belief networks and recurrent neural networks”

Definition of deep learning

- Rebranding of neural networks
  - “Some of the representations are … loosely based on … communication patterns in a nervous system, such as neural coding which attempts to define a relationship between various stimuli and associated neuronal responses in the brain.”
- Competitive results
  - “in computer vision, automatic speech recognition, natural language processing, audio recognition, and bioinformatics”
- Similar hype than with neural networks in the 90s
- Extremely successful especially in vision with n >>> p, but overfitting for moderate n

A statistical view of deep learning



The Spectator
Shakir's Machine Learning Blog
★ About Me
Learning in E
Post Collections
A Statistical View of Deep Learning
3

A Statistical View of Deep Learning

Over the past 6 months, I've taken to writing a series of posts (one each month) on a statistical view of deep learning with two principal motivations in mind. The first was as a ***personal exercise*** to make concrete and to test the limits of the way that I think about, and use deep learning in my every day work. The second, was to highlight important statistical connections and ***implications of deep learning that I do not see being made*** in the popular courses, reviews and books on deep learning, but which are extremely important to keep in mind.

Post Links and Summary

Links to each post with a short summary and as a single PDF are collected here.

1. **Recursive GLMs.** We constructed a view of deep feedforward networks as a natural extension of generalised linear regression formed by recursive application of the the generalised linear form. Maximum likelihood was shown to be the underlying method for parameter learning.

http://blog.shakirm.com/ml-series/a-statistical-view-of-deep-learning/

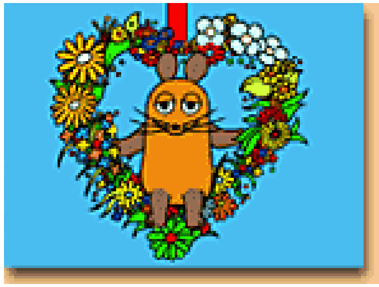
A statistical view of deep learning

- Deep feedforward networks
  - Natural extension of generalized linear regression
  - Recursive application of the generalized linear form, with maximum-likelihood for parameter learning
- Recurrent networks
  - State-space models or dynamical systems
  - Recurrent networks assume that hidden states are deterministic, state-space models have stochastic hidden states
  - Maximum-Likelihood reasoning, innovative new models
- Various forms of statistical regularization implemented

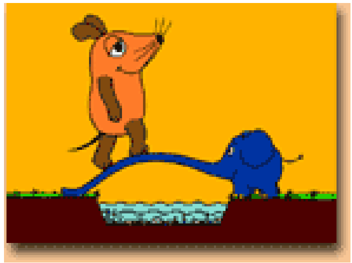
Thank you

- Thank you very much for your attention !

Biology



Computer Science



Statistics