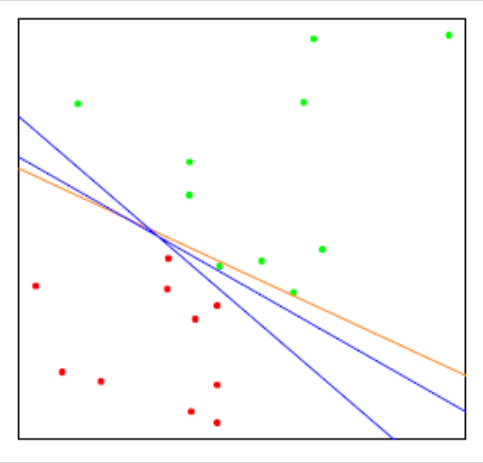


Big Data Analytics

Basic concepts of analyzing very large amounts of data

Dr. – Ing. Morris Riedel
Adjunct Associated Professor
School of Engineering and Natural Sciences, University of Iceland
Research Group Leader, Juelich Supercomputing Centre, Germany

LECTURE BDA5

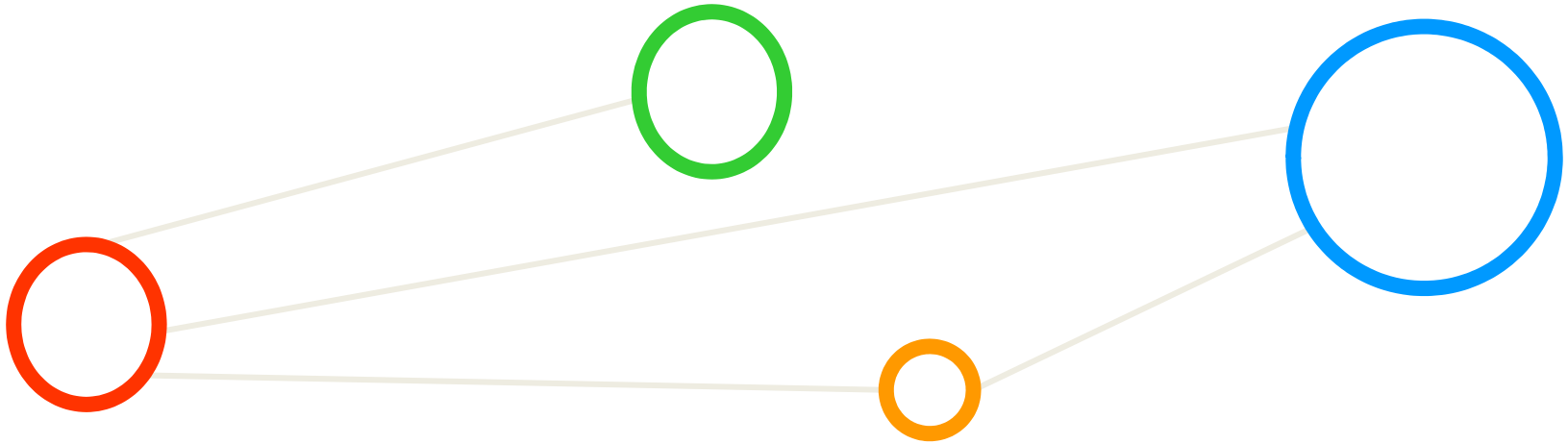
Tools and Techniques

2014-02-08 (v2)
Online Material

UNIVERSITY OF ICELAND
SCHOOL OF ENGINEERING AND NATURAL SCIENCES
FACULTY OF INDUSTRIAL ENGINEERING, MECHANICAL ENGINEERING AND COMPUTER SCIENCE

JÜLICH
FORSCHUNGSZENTRUM

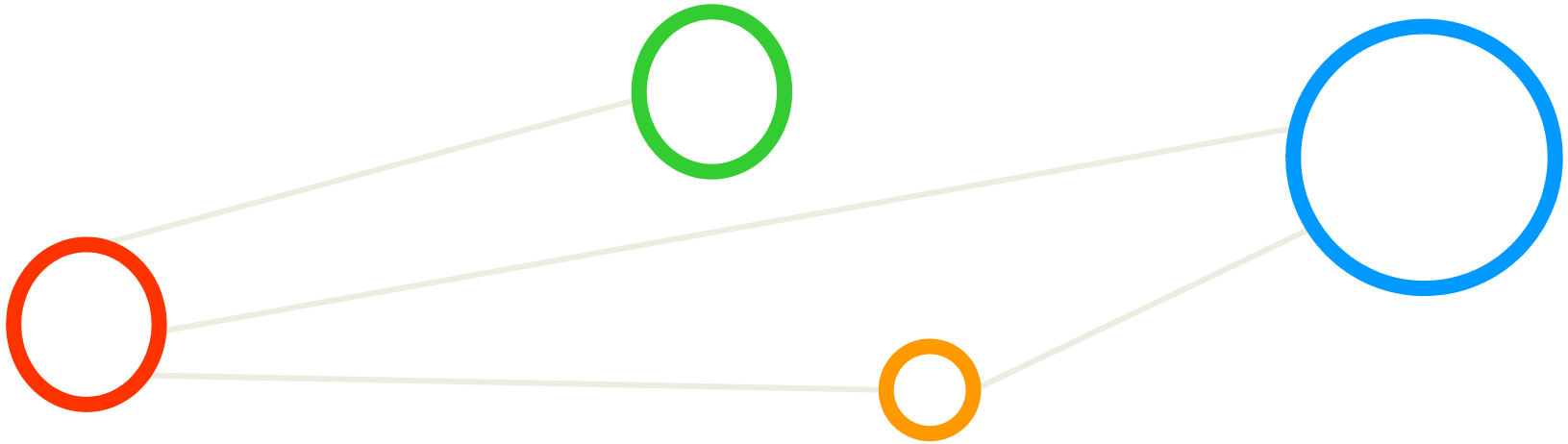
Outline



Outline

- Tools Overview and Selected Use Cases
  - RMPI – R interface to MPI
- Methods
  - Statistical Techniques
  - Data Mining
  - Machine Learning
  - Evolutionary Optimization

RMPI



RMPI – Overview

- **The tool R is a free software environment for statistical computing and graphics**
- **RMPI is an R package and an interface (wrapper) to the Message Passing Interface (MPI)**
- **MPI is a standardized and portable parallel programming model implemented via libraries**

- RMPI ports low level MPI functions into R so that users do not have to know C or Fortran
- MPI libraries supported:
  - LAM-MPI, MPICH(2), and OpenMPI
- Write R programs using certain RMPI functions:
  - Startup and Shutdow: e.g. mpi.spawn.Rslaves([nslaves=#])
  - Cluster Information: e.g. mpi.comm.rank()
  - Use sending/receiving data/functions to send an R object like a number, string, or a list between different slaves or to all slaves (e.g. mpi.bcast.Robj2slave(object))

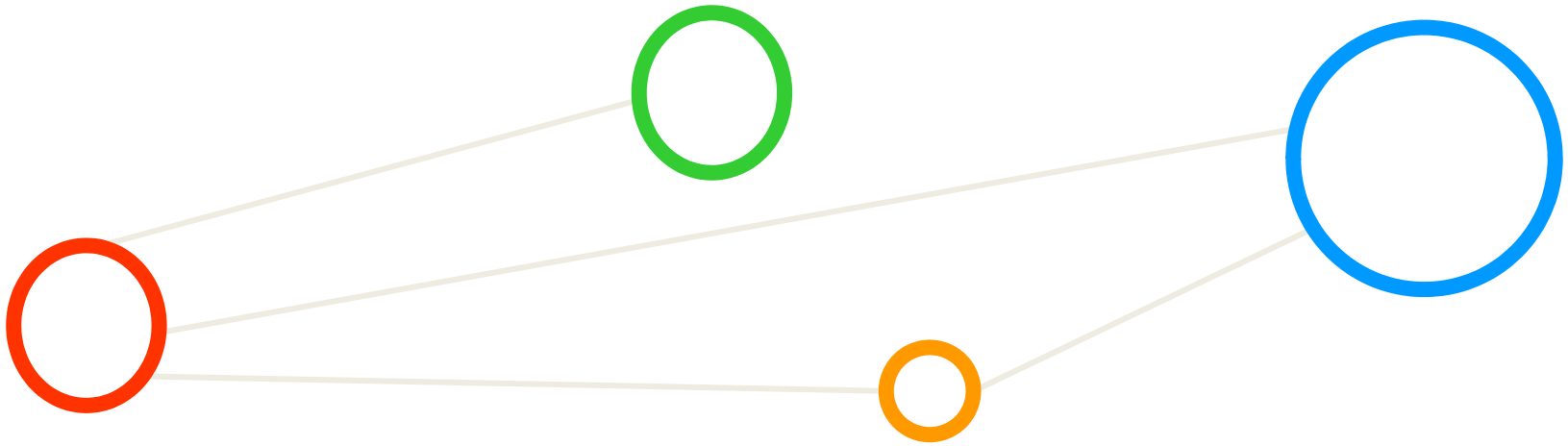
[2] RMPI Web Page

- **Enables inter process communication**

RMPI – Selected Use Cases

- ‘Precision Medicine‘ *[1] Bottomly et al., 2013*
  - Simulations and calculation of P-values and False Discovery Rate (FDR)
  - Parallel computed on Beowulf-style cluster using RMPI R-2.15.1
  - Use of parallel random number generation using L‘Ecuyer‘s method (R rlecuyer package)
  - Plotting and summaries for the simulations with ggplot2 on R-3.0.1

Methods



Statistical Techniques

- **The main task of statistical techniques is to test the likelihood of a given hypothesis**

- Limits
  - Impossible to identify new relationships in data
  - Discovery of new relationships is a very important chance to realize additional unexpected benefits from the large amounts of data
  - Example: data available from High Throughput Experimentation (HTE)

[3] Ohrenberg et al., 2005

Data Mining Techniques

- **Data mining stands for algorithms and techniques which transfer data into information**

- Provides methods that extract rules, patterns, or other structural information from data
- Identify in the data new relationships that were previously undiscovered
- Typical data-mining techniques
  - Clustering, separation methods, decision trees, association rules, etc.
- Limits
  - Data mining does not assess the validity of extracted rules
  - Making data-mining results reliable, it is essential to apply different approaches and check the resulting rules for mutual compatibility

[3] Ohrenberg et al., 2005

Machine Learning Techniques

- **Machine learning techniques are able to learn from data and train a system (model)**

- Wide variety of machine learning algorithms exist
  - Artificial Neural Networks, Support Vector Machines, etc.
- Artificial Neural Networks (ANN)
  - A kind of extremely powerful regression model (rough perspective)
  - Input-output relationships can be generated by ANN as black box model
  - Black box model: function cannot be interpreted physically – the parameters in the ANN have generally no explicit physical meaning
  - ANN need less data then other black-box methods to generate an input – output relationship of comparible quality
    (given a wide range of conditions)

[3] Ohrenberg et al., 2005

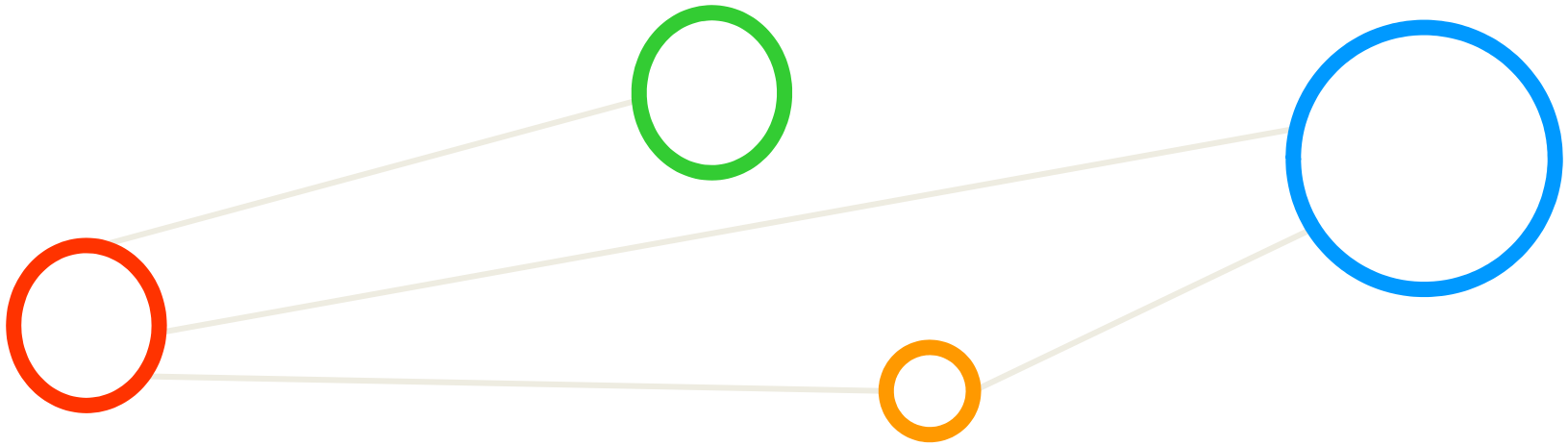
Evolutionary Optimization

- **Evolutionary optimization techniques are characterized by a stochastic approach**

- Advantages
  - Compared to statistical experiment design, evolutionary algorithms are very flexible and adaptable
- Limits
  - In most cases the optimization speed is rather slow

[3] Ohrenberg et al., 2005

Lecture Bibliography



Lecture Bibliography

- [1] Bottomly et al.: Comparison of methods to identify aberrant expression patterns in individual patients: augmenting our toolkit for precision medicine. Genome Medicine 2013 5:103., Online: http://genomemedicine.com/content/pdf/gm509.pdf
- [2] RMPI Web Page, Online: http://www.stats.uwo.ca/faculty/yu/Rmpi/
- [3] Ohrenberg, Arne, et al. "Application of Data Mining and Evolutionary Optimization in Catalyst Discovery and High-Throughput Experimentation–Techniques, Strategies, and Software." *QSAR & Combinatorial Science* 24.1 (2005): 29-37.


services
measurement
within
fusion
new
funding
Policy-based
concepts
device
Analysis
forms
cross-disciplinary
resources
Centre
climate
Computer
expertise
Approaches
construction
Fusion
year
like
Computational
Test
E.g
Computing
disciplines
results
Enable
Cross-Disciplinary
methods
High
big
nodes
PB
community
Environment
Modelling
different
EU
Science
centers
Key
scientific
important
files
model
etc
DKRZ
Use
storage
basis
Infrastructure
international
simulations
ENES
technologies
SMAQ
engineering
References
formats
brain
increasing
via
EUDAT
physical
processing
IPCC
Big
Project
Data
Common
TB
MapReduce
create
computing
Using
Methods
European
databases
GB
Supercomputing
Work
images
codes
analysis
DLCL
Level
stored
Resources
analyze
Large
Summary
data-intensive
Service
Scientific
levels
often
use
external
JSC
computational
Research
Matters
Concrete
performance
research
e.g
science
HPC
large-scale
Juelich
system
datasets
acquisition
compute
Landscape
System
Climate
modelling
AR5
Hardware
distributed
cores
access
hundreds
Nutshell
Services
management
Technologies
Earth
Structure
current
network
DLCLs
Understanding
structures
directory
project
General
Germany
Simulation
Processing
Provide
NASA
Energy
systems
day
Health