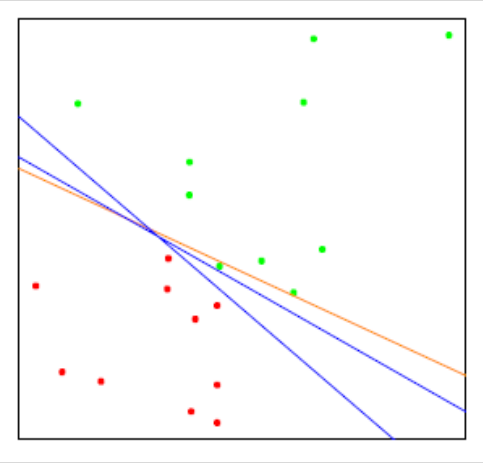


Big Data Analytics

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 BDA4

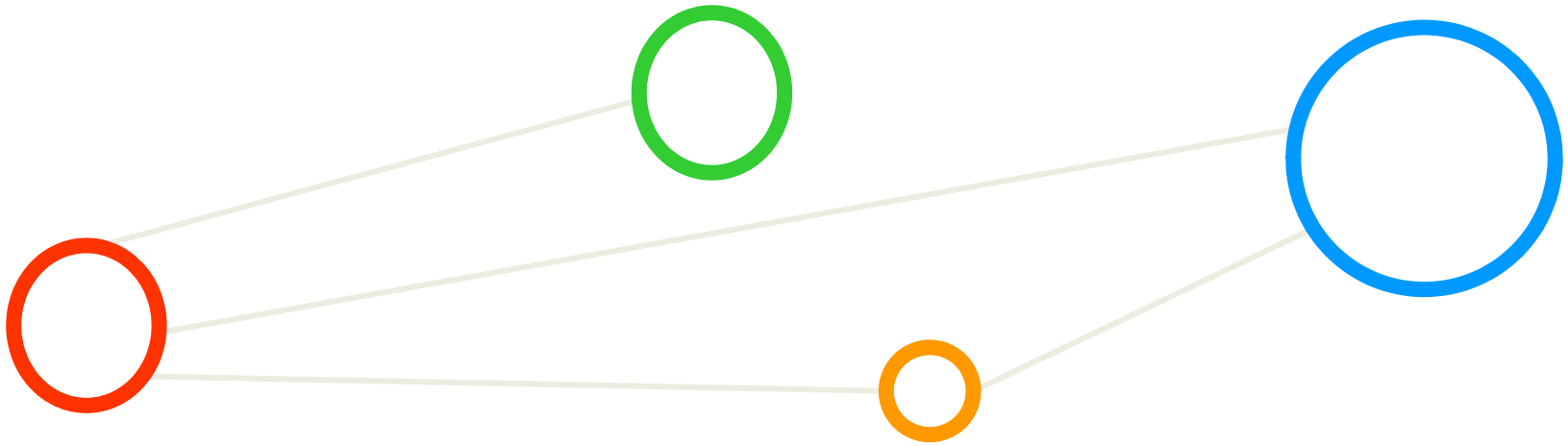
Research Challenges

2014-01-27 (v1)
Online Material

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

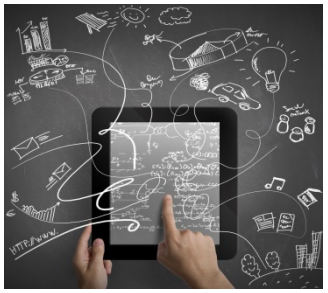
JÜLICH
FORSCHUNGSZENTRUM

Outline

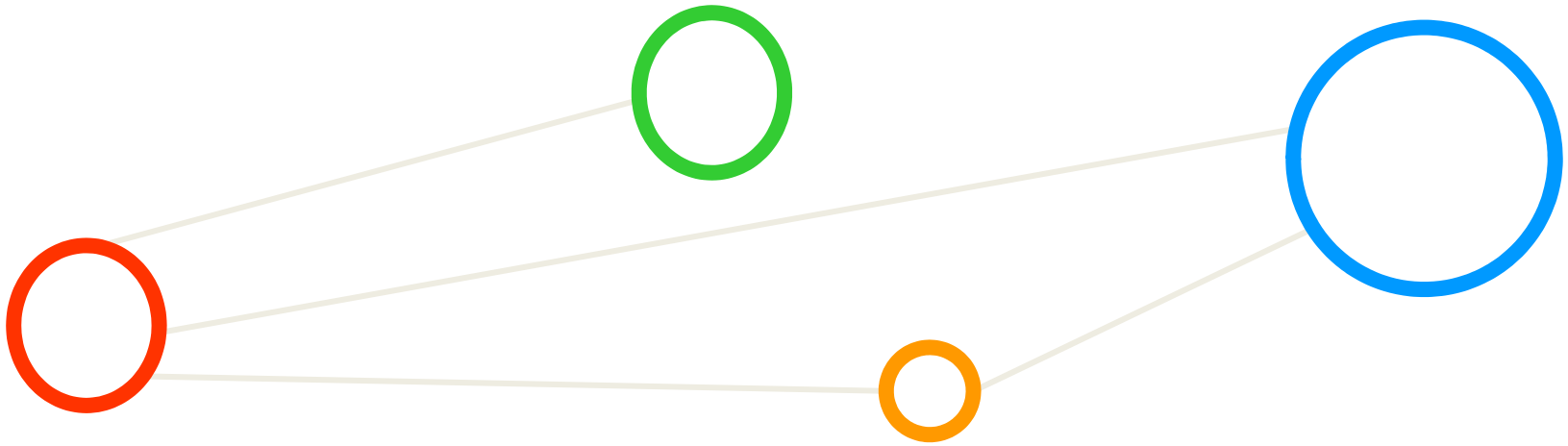


Outline

- Classification
  - RDF Documents and Linked Data



Classification



RDF Documents and Linked Data (1)

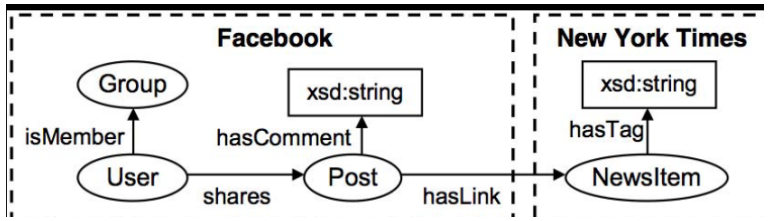
- Data: Resource Description Framework (RDF) Stores
  - Interlinked, physically distributed, and autonomously maintained
  - Community-driven Linked Open Data (LOD) effort allows structured data to be represented using RDF using subject-predicate-object 'RDF triples'
  - Associated query languages such as SPARQL offer the means to store and query large amounts of RDF data
- Goals: Predictive modeling and knowledge discovery
- 'Limited applicability' of existing machine learning techniques
  - Neither desirable nor feasible to have all data in a centralized location
  - Analysis problematic due to access, memory, bandwidth, computational restrictions, privacy/confidentiality constraints

- **Problem statement: Learning predictive models from multiple interlinked RDF data stores**
- **Technique for learning predictive models (e.g. classifiers) from multiple interlinked RDF stores that support only indirect access to data (e.g. via a query interface like SPARQL)**

[1] H. Lin et al., 2013

RDF Documents and Linked Data (2)

- Predictive modelling example
  - Instance of ‘node prediction problem‘:
  - ‘One might want to use data from Facebook and New York Times to predict the interest of a user in belonging to a Facebook group, based on the distribution of tags associated with the New York Times news stories that the user has shared with her social network on Facebook.’



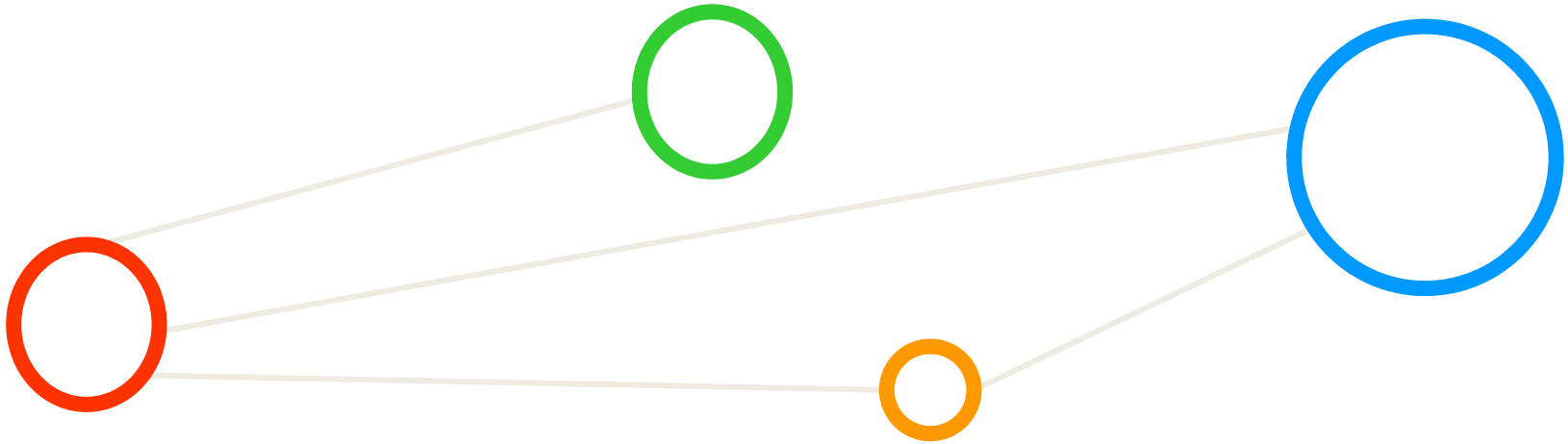
- Research challenges, approaches, and contributions
  - ‘Statistical query based formulations‘ of several learning algorithms
  - ‘Distributed learning framework‘ to learn from interlinked RDF stores
  - Three special cases of ‘RDF data fragmentation‘
  - ‘Matrix reconstruction technique‘ from field of Computerized Tomography (to reduce the amount of information transmitted from remote sources)
  - Experiments with ‘real-world social networking data set‘ ***[1] H. Lin et al., 2013***

RDF Documents and Linked Data (3)

- Learn from RDF data
  1. Approaches that rely on *aggregation* to encode nodes to be classified as 'tuples of attribute values' (instances that can be handled by traditional supervised machine learning algorithms)
  2. Approaches that are based on generative models of data
- Approach 1 – Aggregation
  - Represent each bag of attributes by a single value
  - Apply a suitable aggregation function (e.g., min, max, average for continuous values and mode for discrete values
  - Reduce the data set into a traditional attribute-value data set
  - Each instance is represented by a finite number of attributes
  - Each attribute takes a single value from the set of possible values
- Applying an aggregation scheme to each of the instances can effectively reduce the problem of learning from an RDF data set to the well-studied problem of supervised learning

[1] H. Lin et al., 2013

Lecture Bibliography



Lecture Bibliography

- [1] H. Lin et al., 'Learning Classifiers from Chains of Multiple Interlinked RDF Data Stores', 2013 IEEE International Congress on Big Data, DOI 10.1109/BigData.Congress.2013.22,
  Online: http://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=6597124


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