Big Data IT-forum

2017-05-09

Eirik Thorsnes
Center for Big Data Analysis, Uni Research AS

uni Computing

Center for Big Data Analysis

- Officially started January 2015 in Uni Research Computing
- Promote big data and machine learning
  - in both science and industry
- Focus is on:
  - Operational big data IT-cluster in-house
  - Machine learning
  - Middleware development, and conversion for scientific data
  - Close collaboration with various domains (science and industry)

uni Research

Big Data and Machine Learning

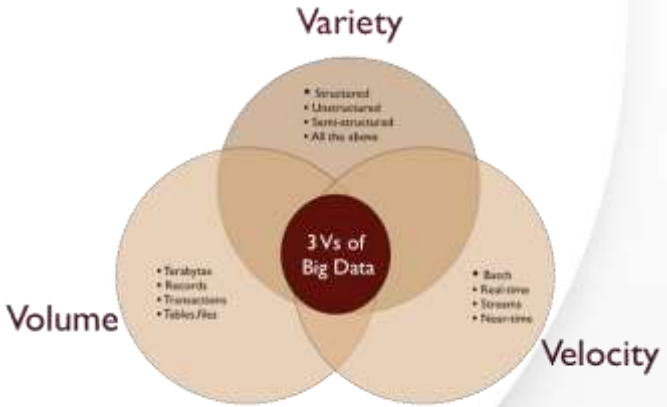
- BD & ML is an approach to understand complex systems
- This approach is made up of several pieces
  - *Complex system (e.g. climate simulation, hydro-power)*
  - *Data flow from the system (e.g. files, sensor readings)*
  - *Big data IT*
  - *Machine learning*
  - *Domain knowledge (e.g. climate science, hydro-power)*
  - *Interaction with the system*
- *The approach is more than its pieces: understanding each piece by itself does not mean to understand the system*

uni Research

Big Data

- Big data comes from web giants like Google, Yahoo, Facebook and Twitter
  - Less used within science

- Big data – not necessarily **big**
  - Complex
  - Streaming
  - Variety of data-types and formats
    - numbers, text, images, video…

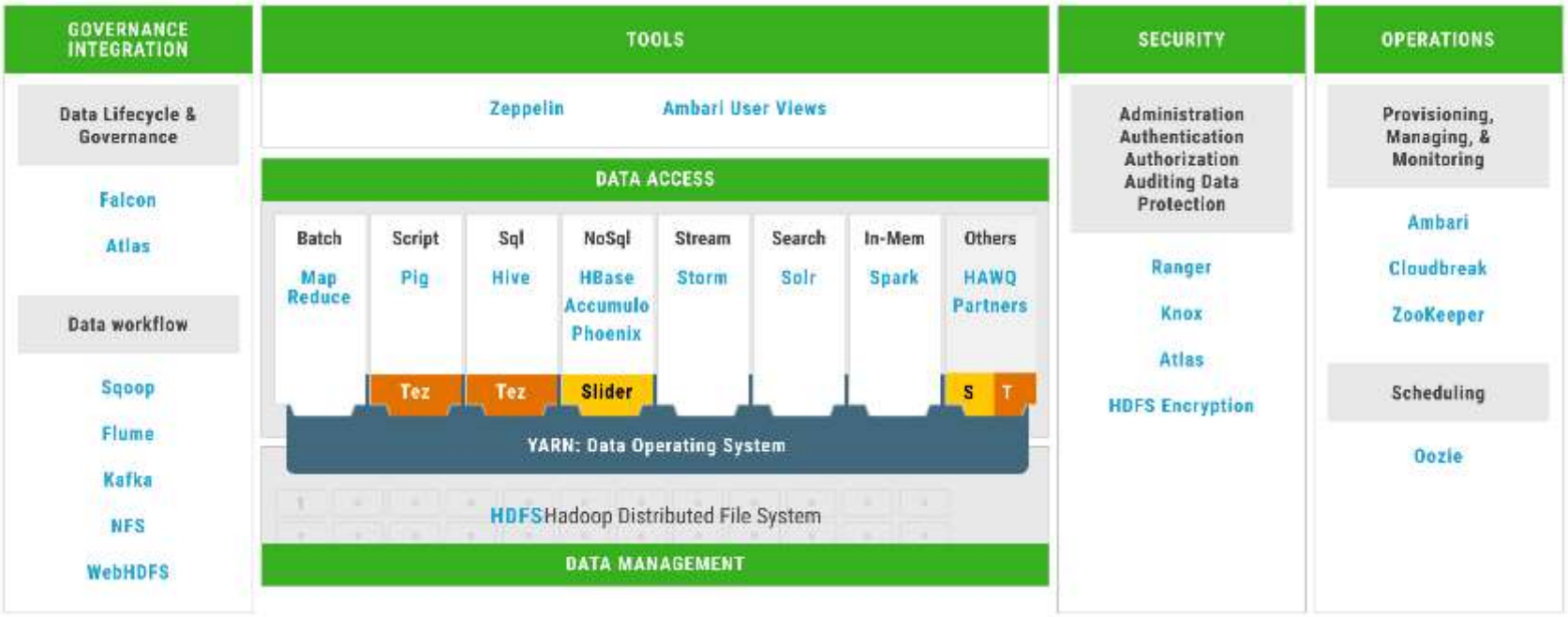


- Big data provides a framework for storage, processing, automation, security, vizualisation/interaction

uni Research

Big Data IT

- Multiple layers create a platform
  - Hardware + OS + deployment/config
  - Hadoop “ecosystem” + glue + application/ML



uniResearch

Big Data – data pipeline

- Data is often “piped” through a chain:
  - Copy into the cluster
  - Format conversion
  - Cleaning
  - Data structuring, de-normalization
  - Metadata and Interface-API for later loading
  - Statistics and machine-learning investigation + analysis
  - Results, REST-API, interaction, vizualization

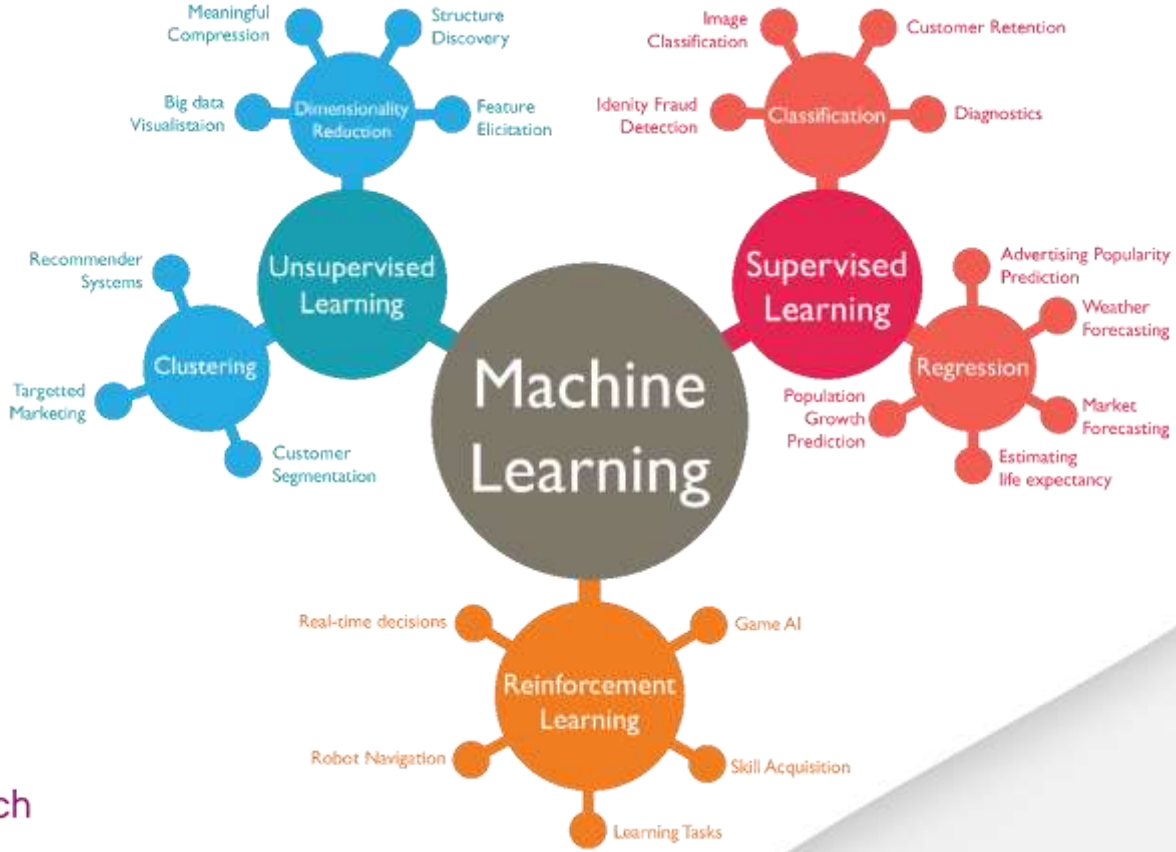
- Streaming data can do this in real-time for e.g. dashboards or decision-support systems

uniResearch

Machine Learning

- Big toolbox of methods and algorithms
- Many of the methods are similar to how humans learn
  - Supervised (learn by examples)
  - Un-supervised (explore structure, group, simplify)

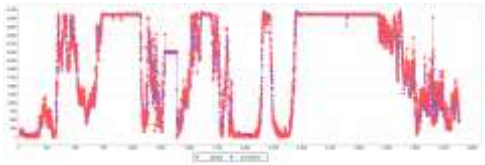
.



uni Research

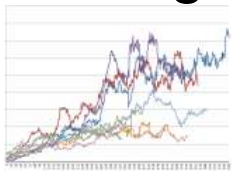
Machine Learning

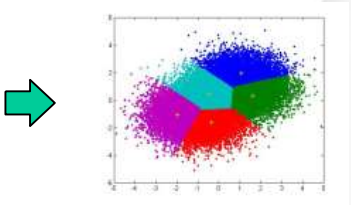
- ML can predict what is going to happen, by knowing what has happened before *(e.g. time-series analysis for wind-turbine production...)*



- ML can discover relationships
  (if this goes up, and this goes down, then we have this situation …)

- ML can group similar things together to give an overview
  (there are five groups of sensor readings…)





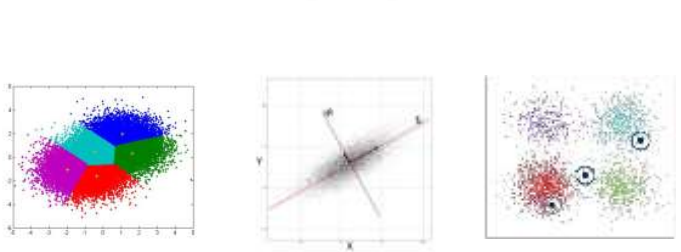
- *Note: Humans can do the same, for a few parameters, ML together with BD can do it for 1000s, and can do it fast.*

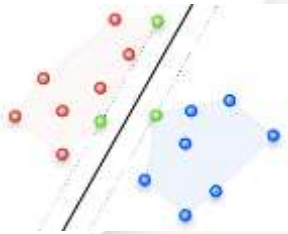
uni Research

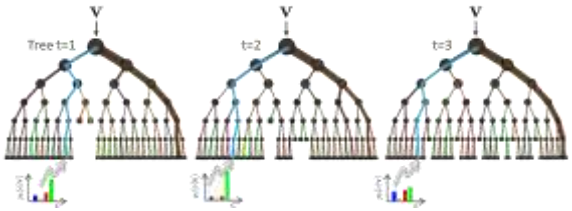
Machine Learning

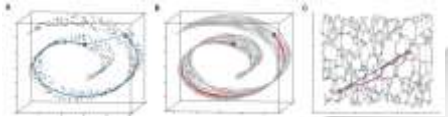
- ML is not a magic box!
- The key is to have a data-scientist to select the best algorithm(s) and parameters for each case

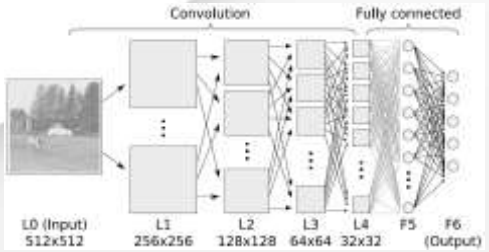
- Artificial Neural Networks
- Convolutional Neural Networks
- “Deep-learning”
- Genetic Algorithms
- k-means
- Multivariate Analysis
- PCA
- Random forests
- Support Vector Machines
- Many more…











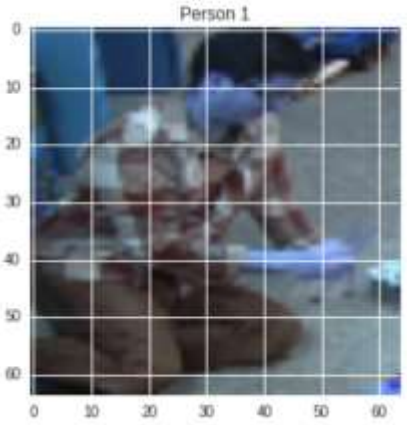
uniResearch

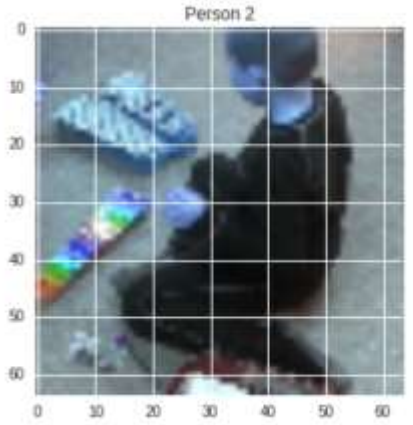
Real-world examples

- Big data + machine learning can be applied to widely different tasks, some of our projects:
  - Prediction of wind-turbine production
  - Identify salmon in underwater video
  - Predict fish-species for fishermen – given time and location
  - Optimize energy usage for ship operations
  - Real time object detection – machine vision
  - Multi-dimensional model reduction – climate change model optimization

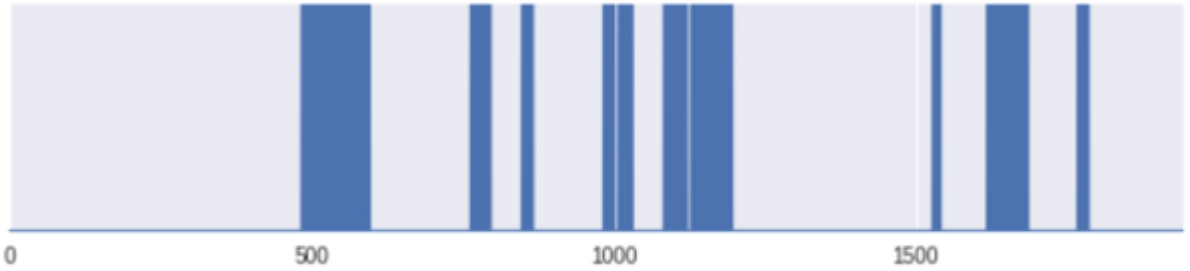
uni Research

Music Therapy session evaluation: person identification, tracing, evaluating interactions





Marked frames with interactions



uni Research

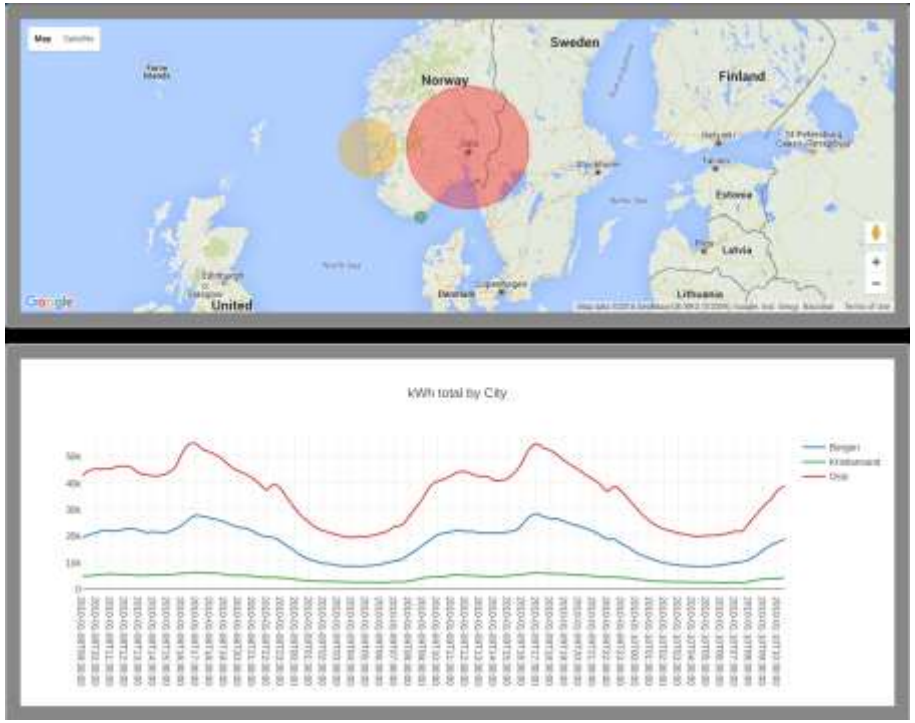
Video Processing

- Real-time traffic monitoring (BT Webcam, Danmarksplass)
- Methods
  - Streaming Analysis
  - Edge detection
  - Background subtraction
  - Object classification & tracking
- Object classification
  - Haar cascades
  - Artificial Neural Networks
  - Deep Learning

uni Research

Power consumption monitoring

- Power consumption aggregates by cities for 2 days period

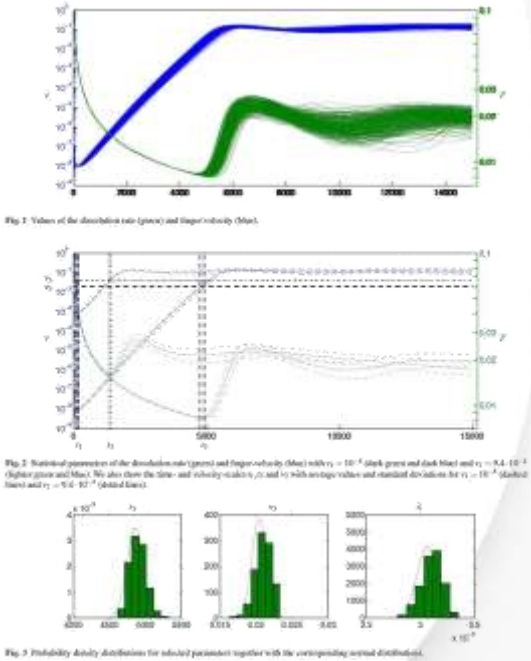


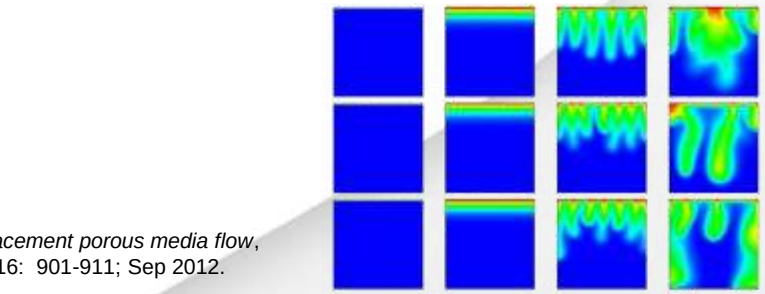
  - Carried out in real-time (200ms, …)
  - Simulated using real data from Ireland

uni Research

Dimension Reduction

- Nonlinear parabolic PDE (porous media)
  - Simulation with about 10^6 DOF
  - Can be described by linear 41 parameters
  - Or by 8-parameters (non-linear)
- That means: The system is complex looking at time development. But is it simple, looking at the structure
- This allows great simplification of the highly complex dynamics





On the time scales of nonlinear instability in miscible displacement porous media flow, M.T. Elenius, K. Johannsen, Computational Geosciences 16: 901-911; Sep 2012.

uni Research

Summary, our experience

- Important to include all aspects
  - Big data engine, IT, operations, machine learning, visualization, interface
- Getting the data into the system takes much more time than you think
- Go with the “big data way”
  - Don’t just replicate an existing code/system/structure
- Big data is interdisciplinary – in itself
  - Connection with application domain (customer) adds to that
  - Team effort also on management level

uni Research

Thank you

for your attention

uni Computing

Extra slides

uni Research

jupyter PySpark-SPD-WindRose Last Checkpoint: 05/03/2016 (autosaved)

File Edit View Insert Cell Kernel Help Python 2

Code CellToolbar

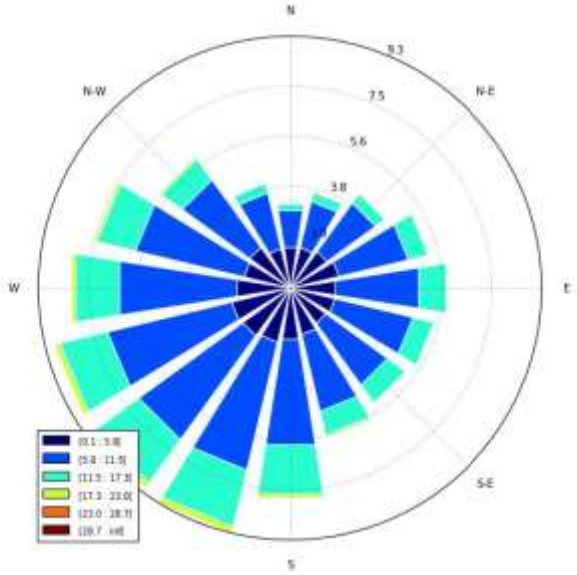
```
start = datetime.datetime.now()

allSPDDF = hiveContext.sql("SELECT spd, dir FROM windspeed")
allSPDPanda = allSPDDF.toPandas()
ws = allSPDPanda['spd'].values
wd = allSPDPanda['dir'].values

ax = WindroseAxes.from_ax()
#ax.contourf(wd, ws, bins=np.arange(0, 8, 1), cmap=cm.hot)
ax.bar(wd, ws, normed=True, opening=0.8, edgecolor='white')
ax.set_legend()

print 'Execution took %s' % (datetime.datetime.now() - start)
```

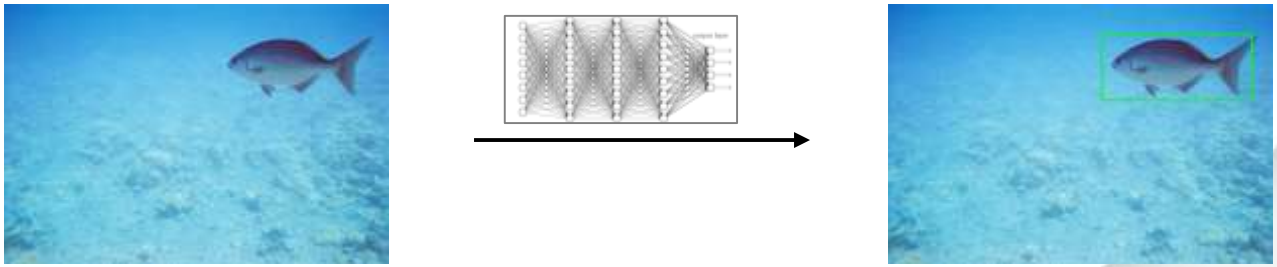
```
Execution took 0:01:11.171768
```



uniResearch

Monitoring of Fish

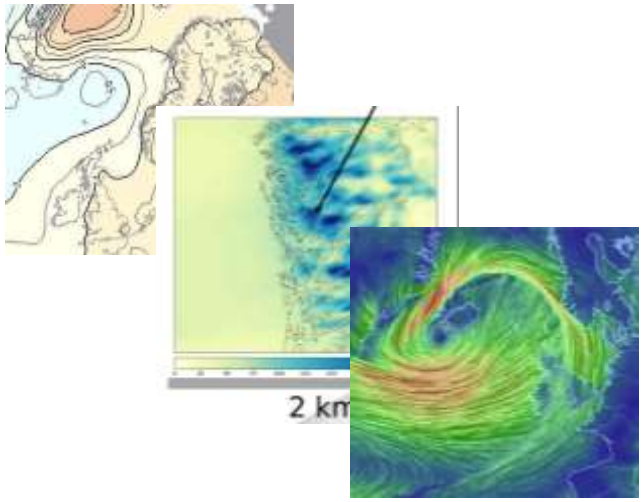
- **Objective:** monitor fish populations in reservoirs and rivers
- **Domain knowledge:** reservoir, fish biology, ecosystem, biodiversity, EU Water Framework Directive
- **Data:** field observations, PIT tagging, scuba diving, sensors, swim-through video recording, trap net fishing
- **Big data, machine learning:** video analysis, time series analysis, predictive modeling, classification, data integration

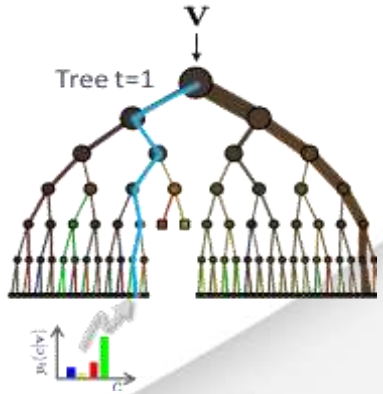


uni Research

Climate and Precipitation Forecast

- **Objective:** weather-, climate and precipitation forecast on time scales from weeks to years available for improved operations planning
- **Domain knowledge:** weather- and climate prediction
- **Data:** Global Forecast System (US), local times series measurements
- **Big data, machine learning:** time series analysis, predictive analysis, expert system for prediction validity

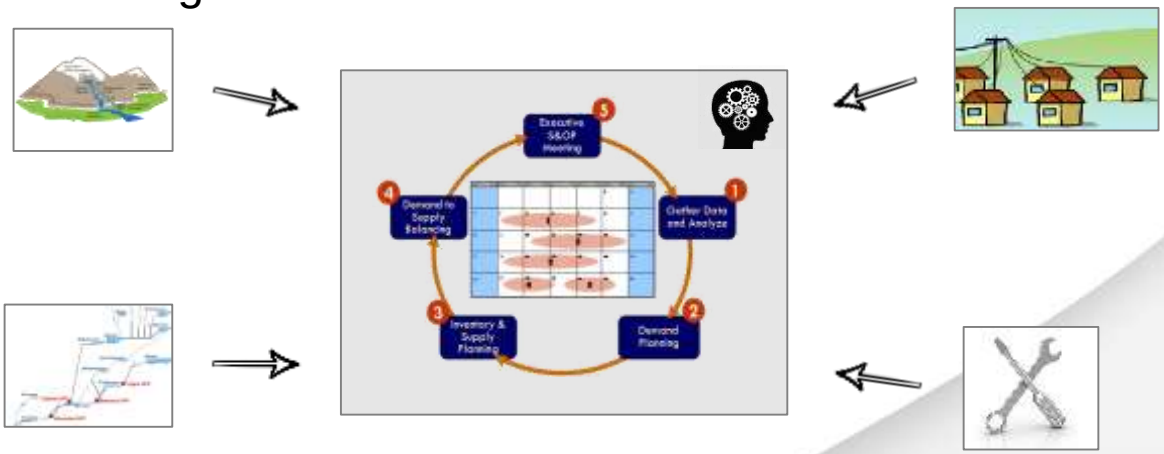




uniResearch

Short Term Operations Planning

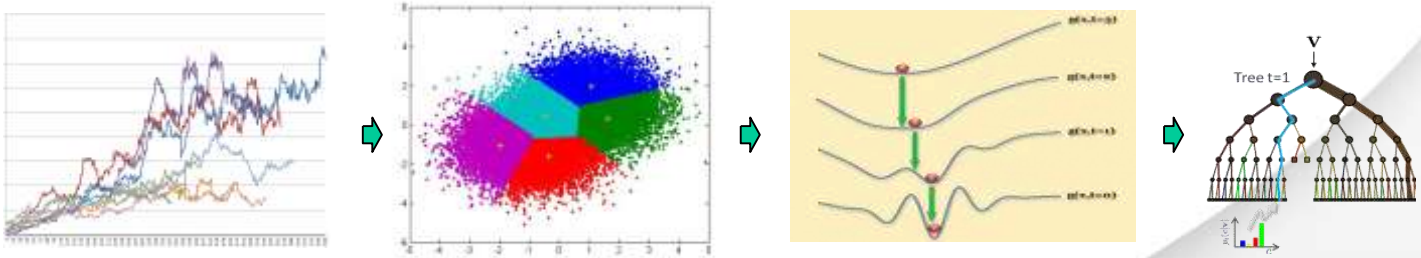
- **Objective:** decision support for optimal operations planning with a time horizon of weeks up to a year
- **Domain knowledge:** from various levels of power grid operations
- **Data:** SCADA, energy consumption, run-off prediction, scheduled maintenance, system upgrades, marketing strategy, policies
- **Big data, machine learning:** optimization, decision trees, classification, Bayesian modeling



uniResearch

Real-Time Operations

- **Objective:** carry out real-time analysis of large amounts of operational data, real-time decision making and/or decision support
- **Domain knowledge:** various levels of power grid operations, smart grid
- **Data:** SCADA, information about scheduled and unscheduled maintenance, system updates, real-time and prediction of energy consumption, run-off prediction, alternative energy sources, a.o.
- **Big data, machine learning:** real-time analysis, classification, optimization, expert system



uni Research