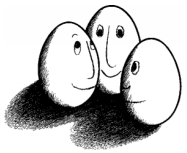
technische universität dortmund



SFB 876 - Providing Information by Resource-Constrained Data Analysis

Big Data Analytics in Astrophysics

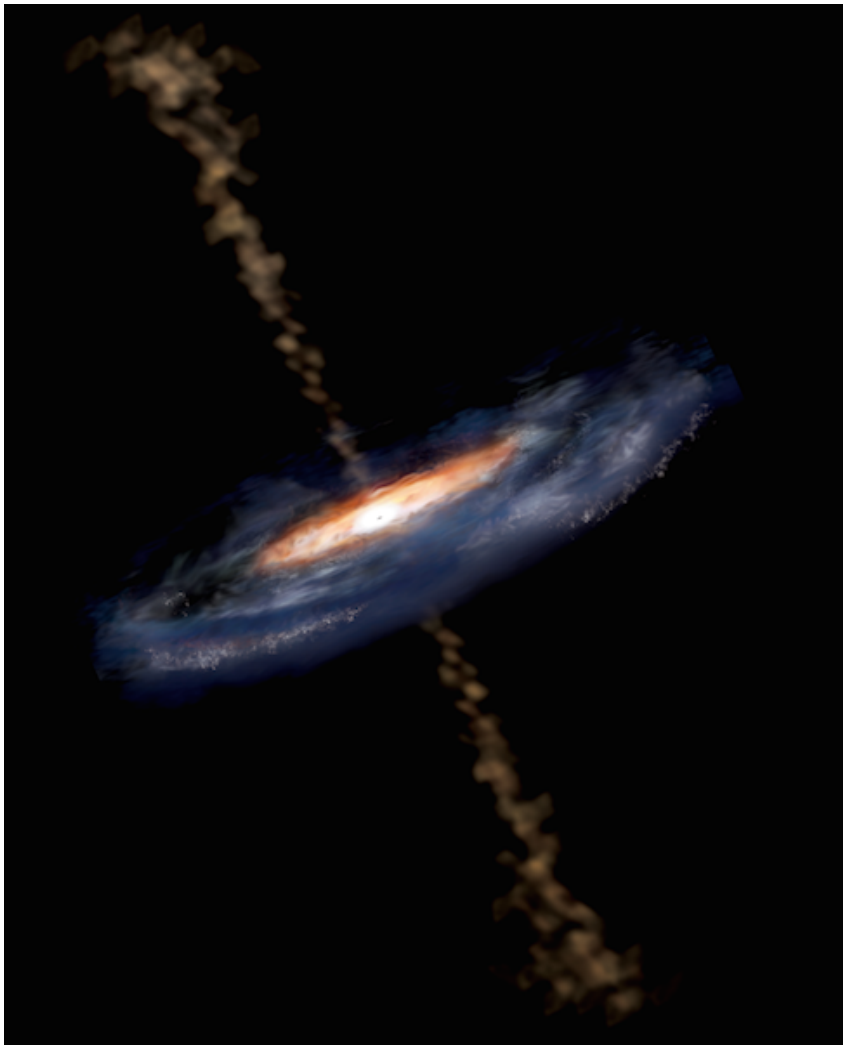
Katharina Morik, Artificial Intelligence,
TU Dortmund

Overview

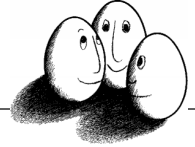
- Short introduction to the Collaborative Research Center SFB 876
- Tools for data analysis, streaming data
- Offline Data Analysis
  - IceCube
- Online Data Analysis
  - Magic, FACT

➢ Science today is based on data.

➢ Data analysis is intrinsically tied to the scientific process.



Active Galactic Nuclei



Collaborative Research Center SFB 876 (2011 -) Providing Information by Resource-Constrained Data Analysis

14 projects
20 professors
50 Ph D students

Small devices, FPGA, energy-restricted devices ...

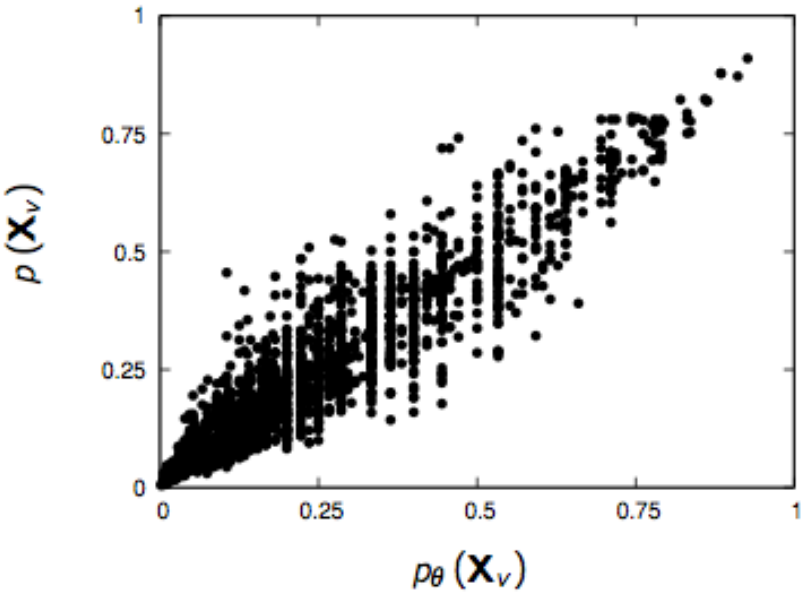
Resource-aware algorithms for big data, streaming data, new computing architectures ...

Computer science theory for data analysis under resource constraints!

New Algorithms for Resource Constrained Data Analysis

- Probabilistic graphical models that use only integer computing – data analysis at very restricted devices.
- FPGA executing models learned by the SVM or deep learning – model application with very low energy consumption.
- Real-time processing of data streams – event detection at large scale.
- ➢ Bringing analysis close to the instruments.

- Some true probabilities (y axis) cannot be expressed by the integer approximation (x axis).



SFB 876: Project C3

- High Energy Astrophysical Phenomena
  - Neutrinos
  - Gamma rays
- IceCube Collaboration
- Magic I, II, FACT

Prof. Dr. Dr. Wolfgang Rhode
Prof. Dr. Katharina Morik
Dr. Tim Ruhe

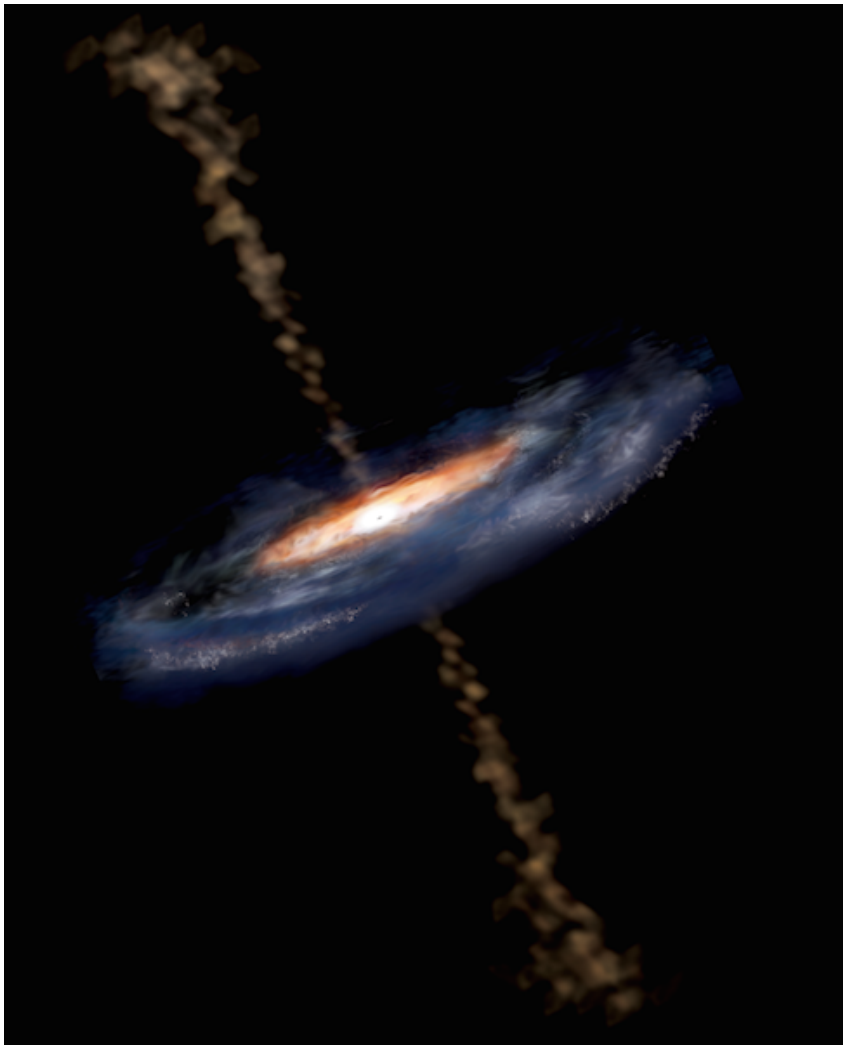
- We are looking for the needle in the haystack: neutrinos and gamma rays are dominated by other particles.
  - Data analysis tool RapidMiner
  - Feature selection MRMR
  - Streams framework for real-time processing

Overview

- Short introduction to the Collaborative Research Center SFB 876
- Tools for data analysis, streaming data
- Offline Data Analysis
  - IceCube
- Online Data Analysis
  - Magic, FACT

➢ Science today is based on data.

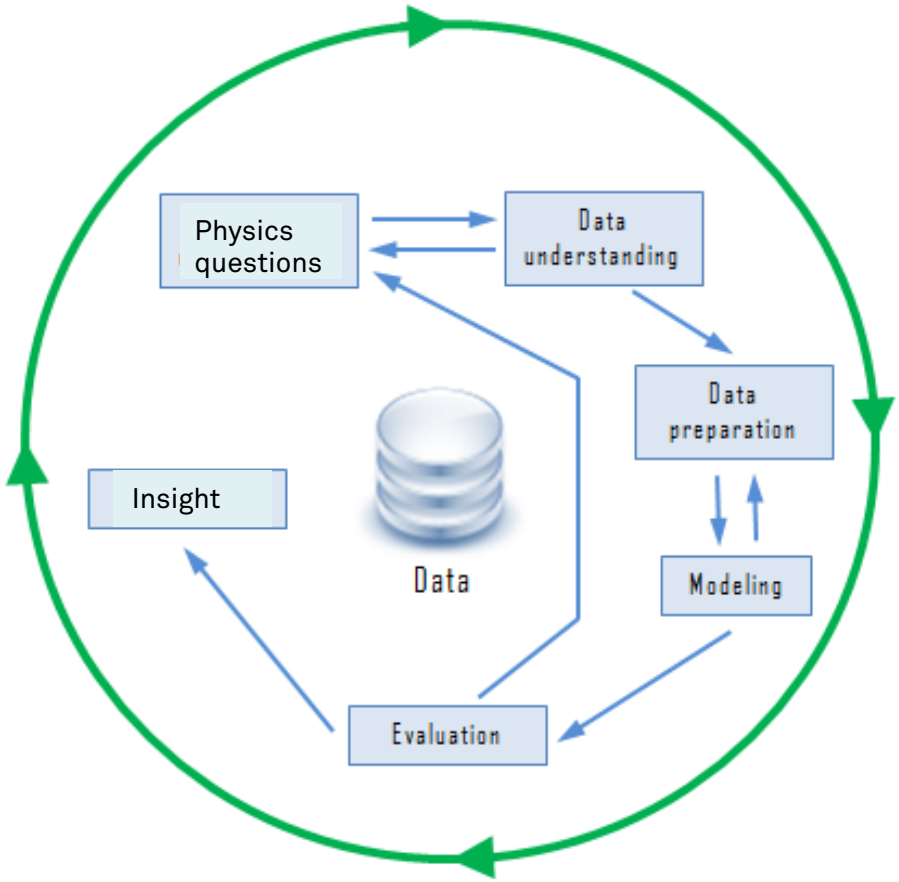
➢ Data analysis is intrinsically tied to the scientific process.

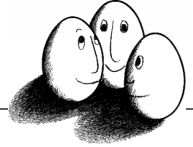


Active Galactic Nuclei

The Data Analysis Cycle

- Data Analysis is not just one step in the overall scientific investigation.
- Data Analysis accompanies the scientific investigation.
- Data Analytics is interdisciplinary in nature:
  - Physics
  - Statistics
  - Data management
  - Machine learning
    - Software/Algorithm engineering
    - Complexity Theory





Data Analysis Tools - Requirements

- Coverage:
  Support the overall cycle
- Amenability:
  Easy to use for all involved scientists
- Rapidity:
  Fast creation of analysis process
- Comparability:
  international scientific terminology
- Reproducibility:
  storing the analysis process with all parameters in a small format.

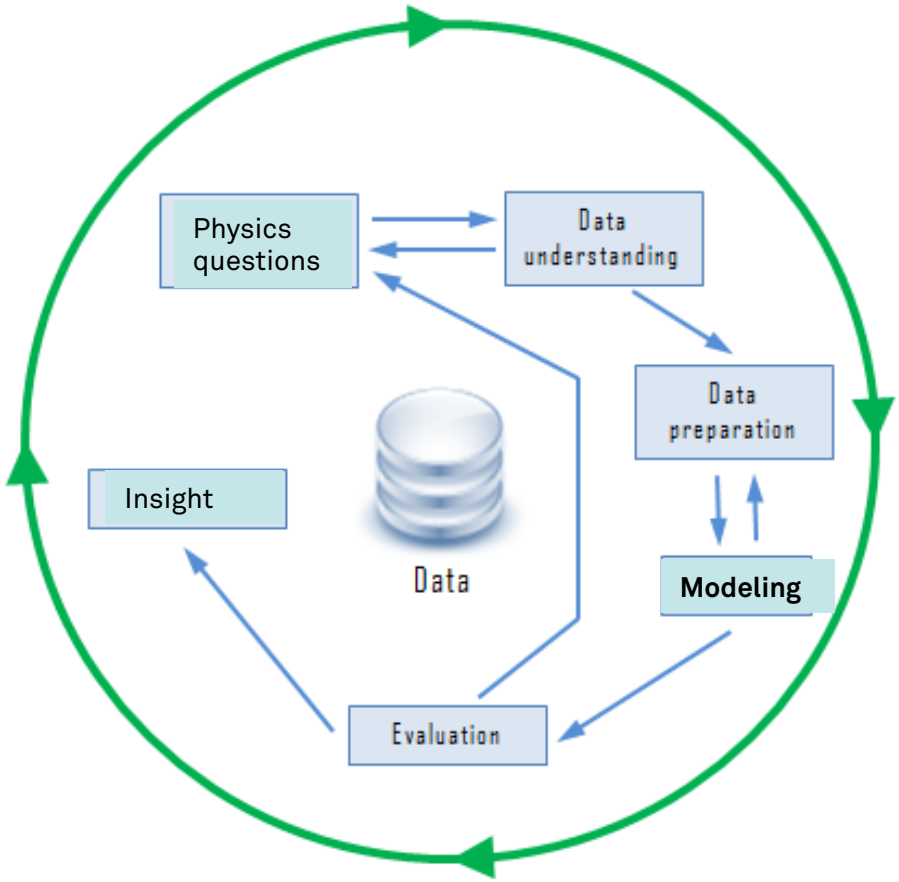
- Not just the modeling step.
- Not just for statisticians, computer scientists, physicists – but for all of them.
- Not a bunch of programs/ libraries to be used in programming.
- Not a new name with a little variance for known methods that have a theoretic basis.
- Not asking the programmer whether he remembers which parameters succeeded.

RapidMiner

- Coverage:
  Support the overall cycle
- Amenability:
  Easy to use for all involved scientists
- Rapidity:
  Fast creation of analysis process
- Comparability:
  international scientific terminology
- Reproducibility:
  storing the analysis process with all parameters in a small format, re-run, exchange.

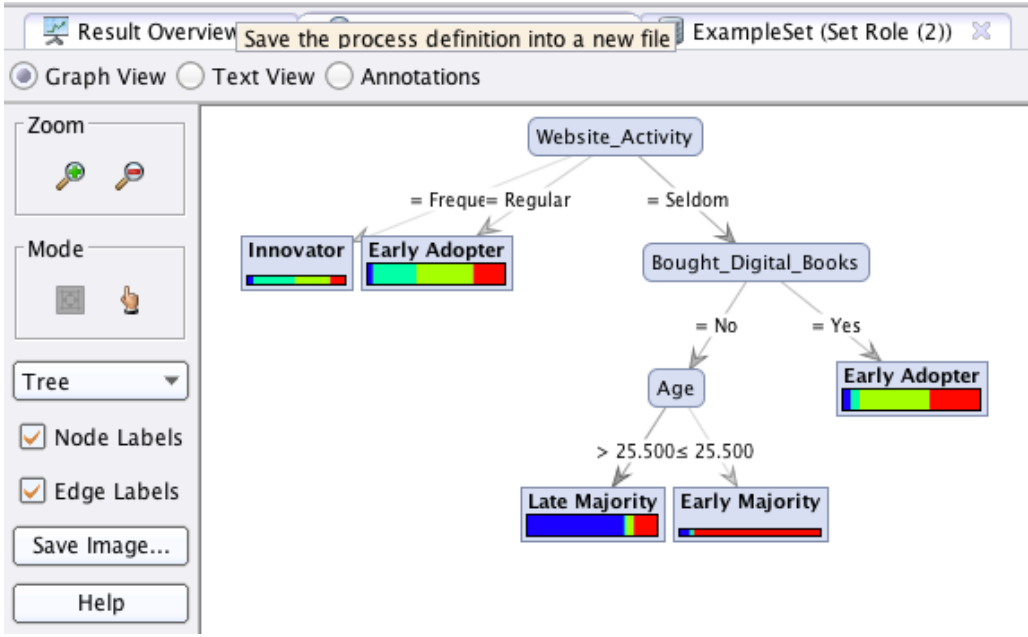
Modeling tasks classification and regression

- Given observations x with labels y $\{(x_1,y_1),...(x_n,y_n)\}$
  - with binominal y (classification)
  - with real-valued y (regression)
- Find f(x)=y such that the error is minimized.
- RapidMiner offers 167 learning algorithms for classification and regression.
- Algorithm should be robust, scalable, parallel!
- Does algorithm implement the formula or approximate it, with which bounds?



Types of Models

- Lazy Modeling, Local Models
  - K-NeirestNeighbors
- Additive Models
  - Decision Trees
- Linear Models
  - Linear Regression
  - Support Vector Machine
- Bayesian Models



Ensemble Methods – here: for decision trees

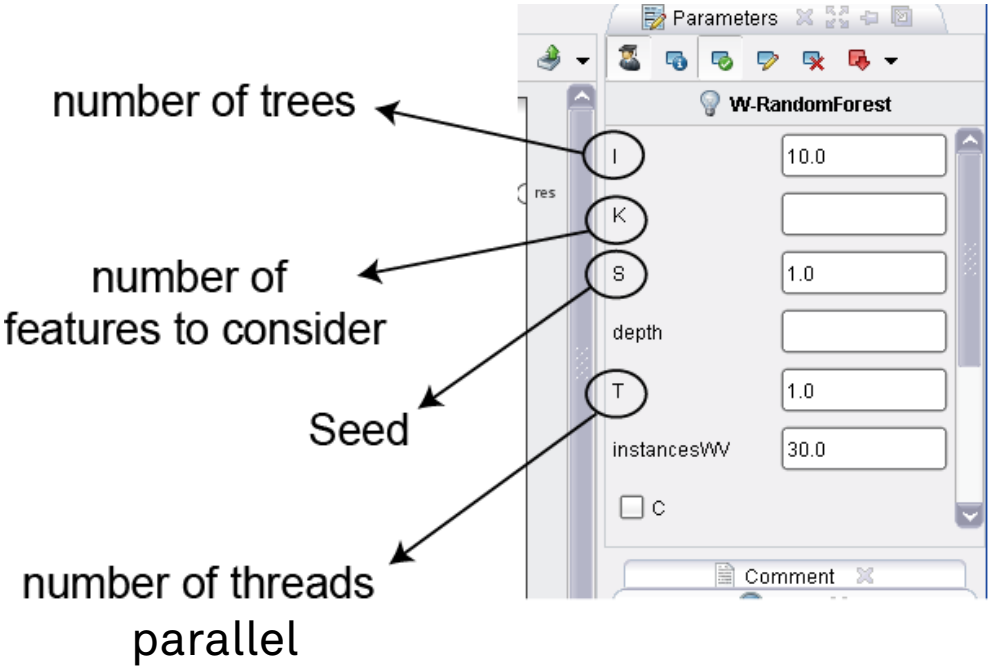
- Ensemble: Take many models and decide according to the majority vote!
- Algorithm Training:
  For l decision trees (parallel):

(1) Take a sample from the data

(2) Take a subset of features

(3) Choose the best feature according to minimal entropy
  - Split according to feature
  - If purity not ok, goto (3).

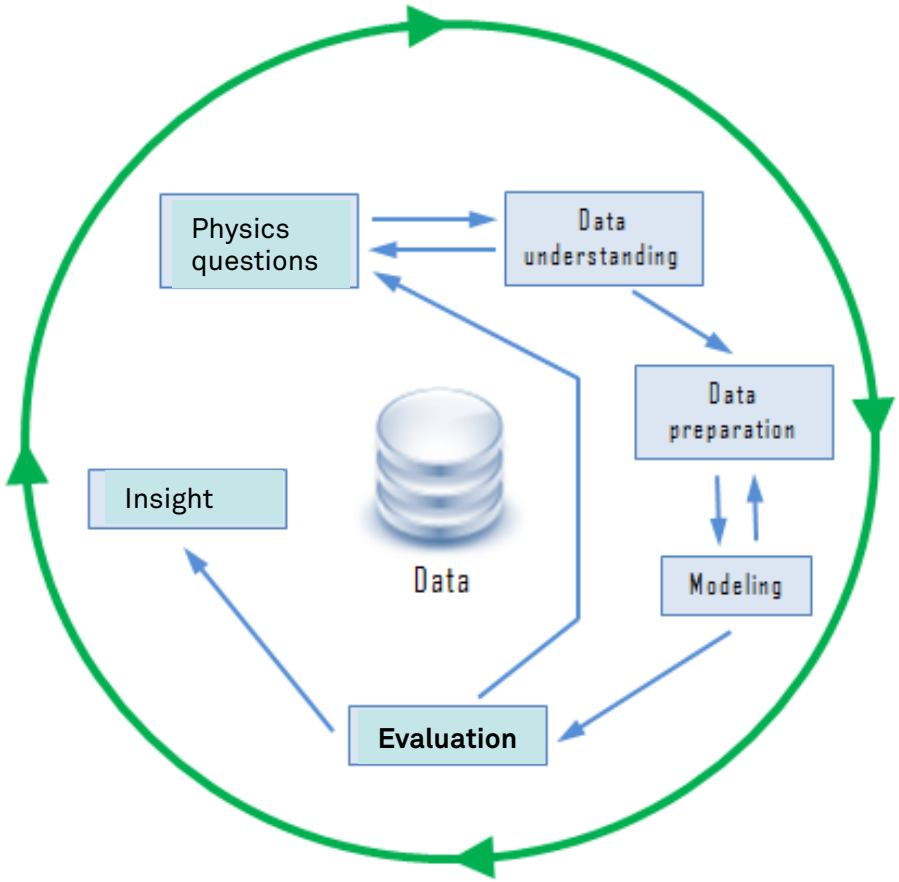
- Weka Random Forest in RapidMiner



Breiman 2001, Machine Learning Journal
Wager 2014, asymptotic bounds, arxiv

Evaluation

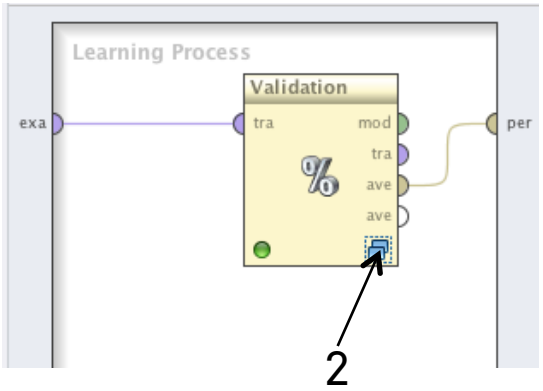
- Training and testing on different samples is necessary in order to estimate the true error.
  - Testing on the same set would overestimate the quality of the model.
- Best test: leave one out requires for n observations n runs of modeling. This is not efficient.
- Crossvalidation: split into m partitions, use m-1 for training and test on the unused one. Output the average of all tests. Fair.



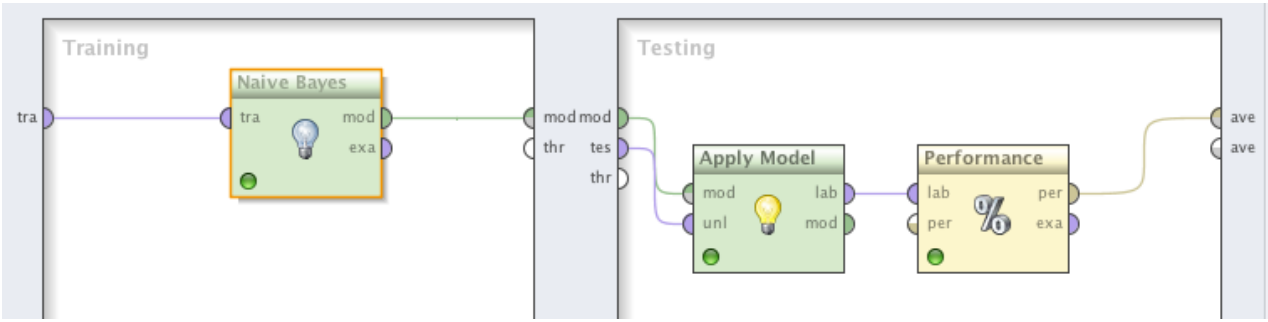
Cross validation in RapidMiner

- Your measurement is meaningless without knowledge of the error.
  Walter Levin

1



2

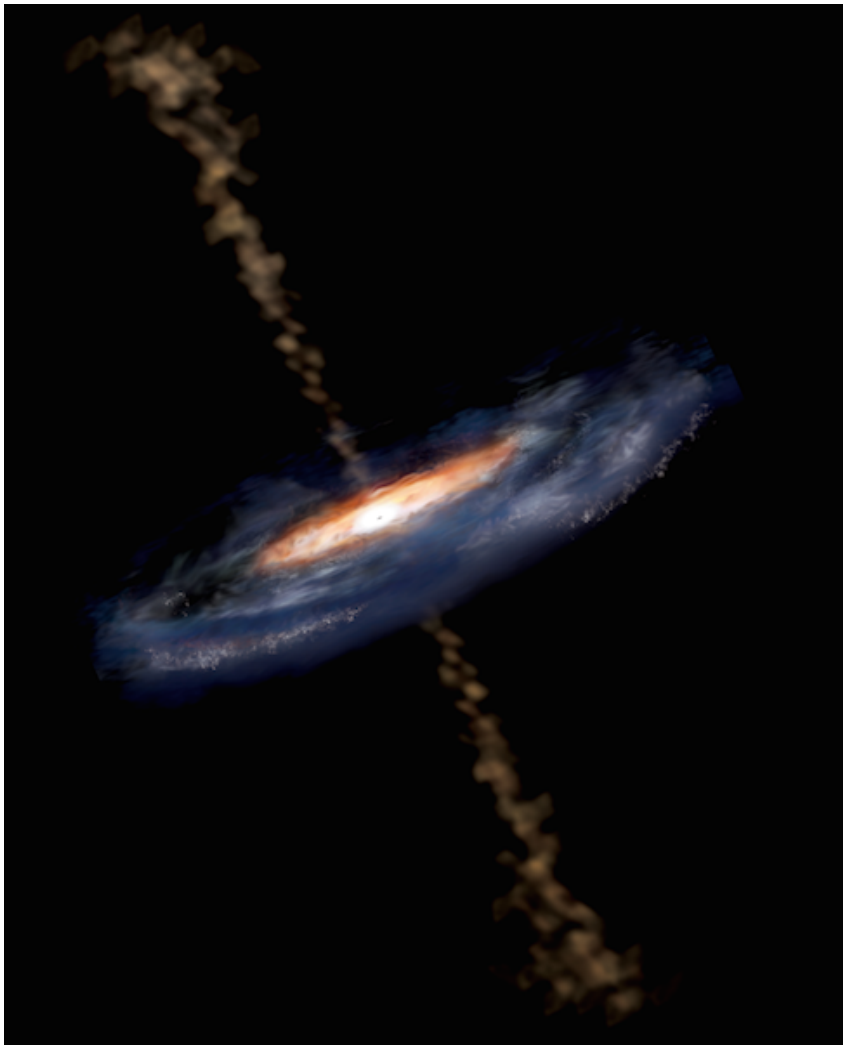


Overview

- Short introduction to the Collaborative Research Center SFB 876
- Tools for data analysis, streaming data
- Offline Data Analysis
  - IceCube
- Online Data Analysis
  - Magic, FACT

➢ Science today is based on data.

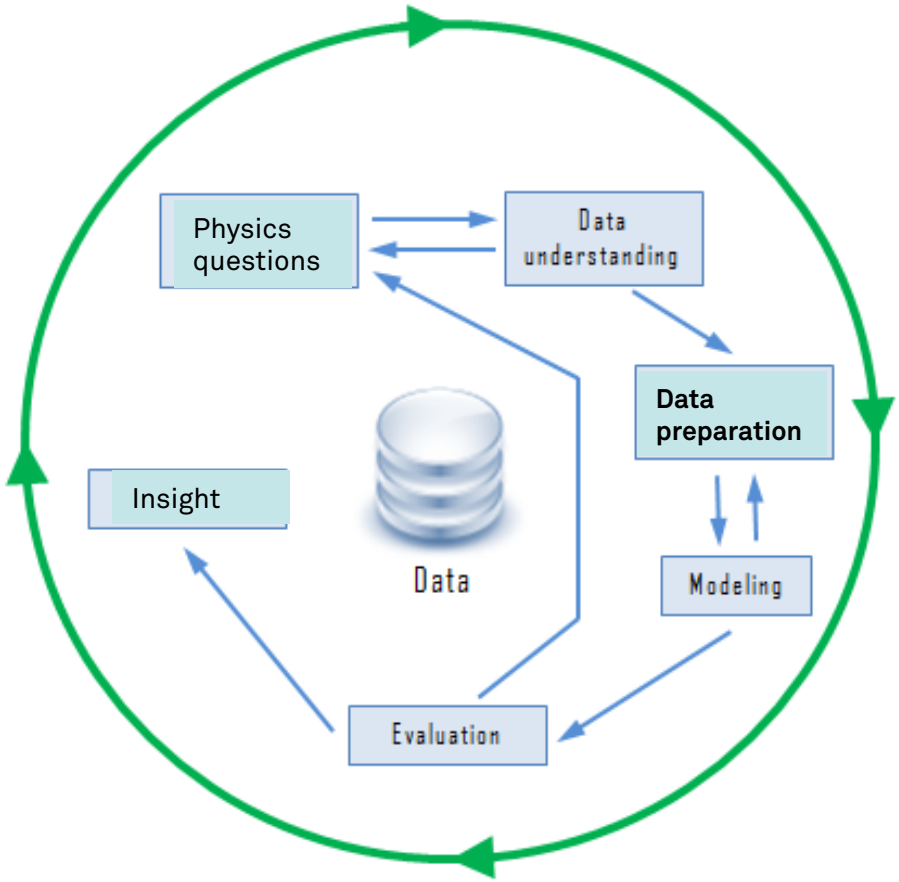
➢ Data analysis is intrinsically tied to the scientific process.



Active Galactic Nuclei

Data preparation – Feature selection

- Too many features hide the true pattern and slow down modeling.
- Redundant features may lead to wrong results.
  - Wrong: Two features with the same meaning – two times an impact.
  - Right: Each feature should have half an impact.
- Quality of a feature set is given by the performance of learning using the feature set.



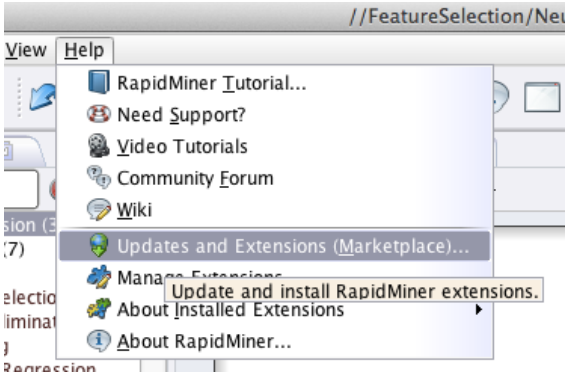
Q -- MRMR

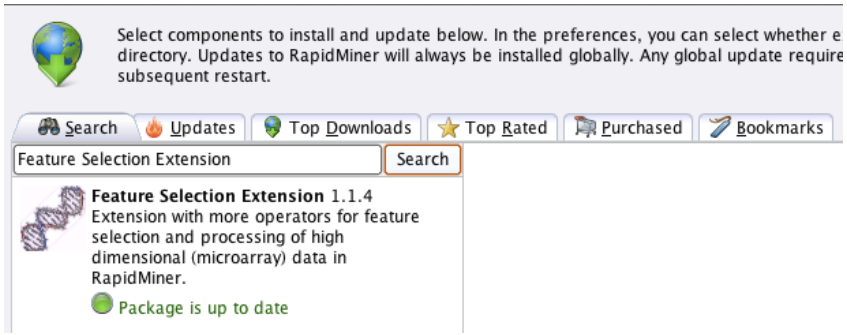
- Minimum Redundancy Maximum Relevance (MRMR)
  - Start with empty feature set.
  - Add the one with lowest redundancy to already chosen features D(x‘,x) and highest relevance w.r.t. label R(x,y)

$$Q = R(x,y) - \frac{1}{j} \sum_{x' in F_j} D(x',x)$$

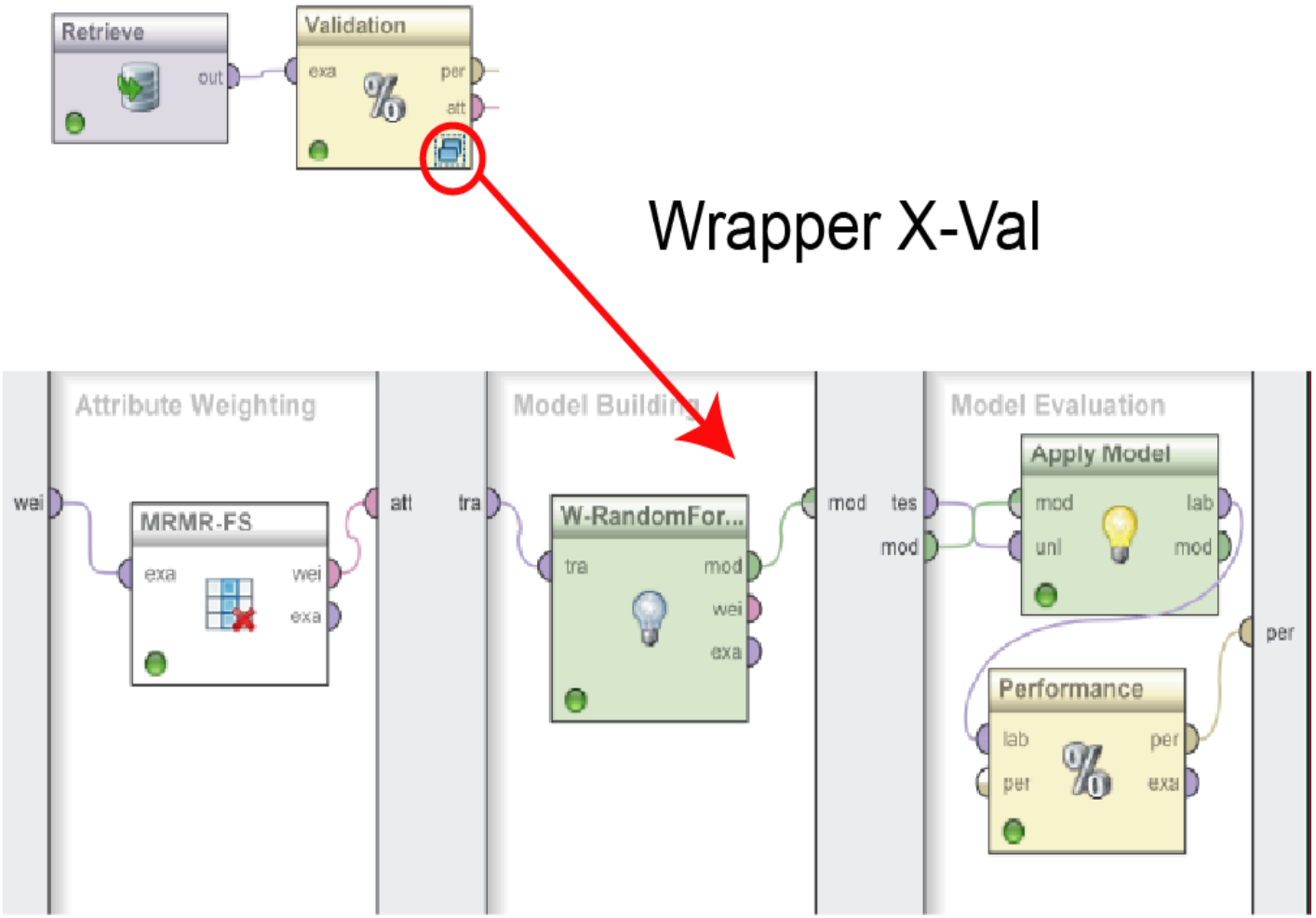
Mark Hall „Correlation Based Feature Selection“ 1999

- Efficient implementation by Benjamin Schowe in RapidMiner‘s Feature Selection Extension



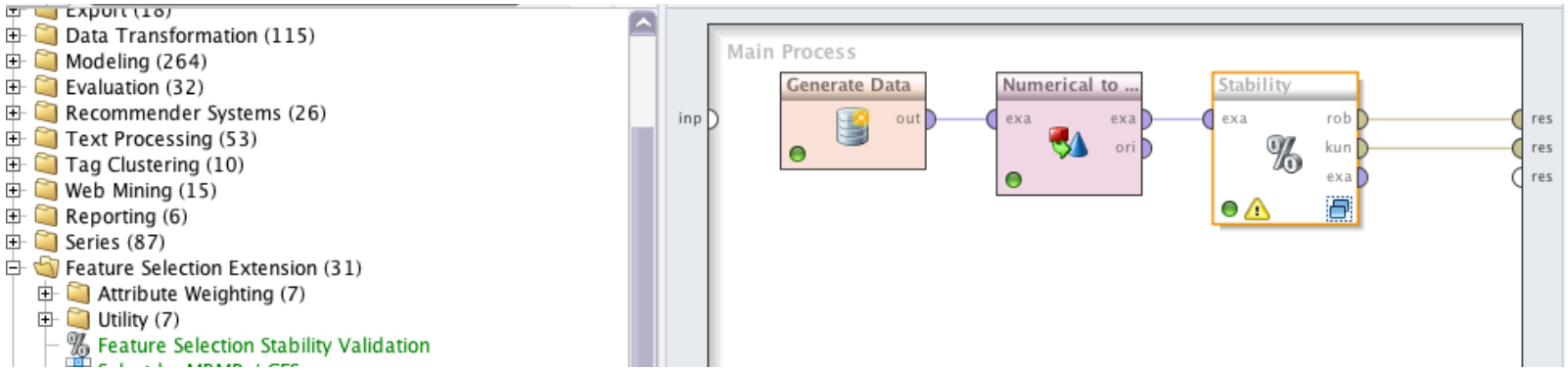


Feature selection in RapidMiner



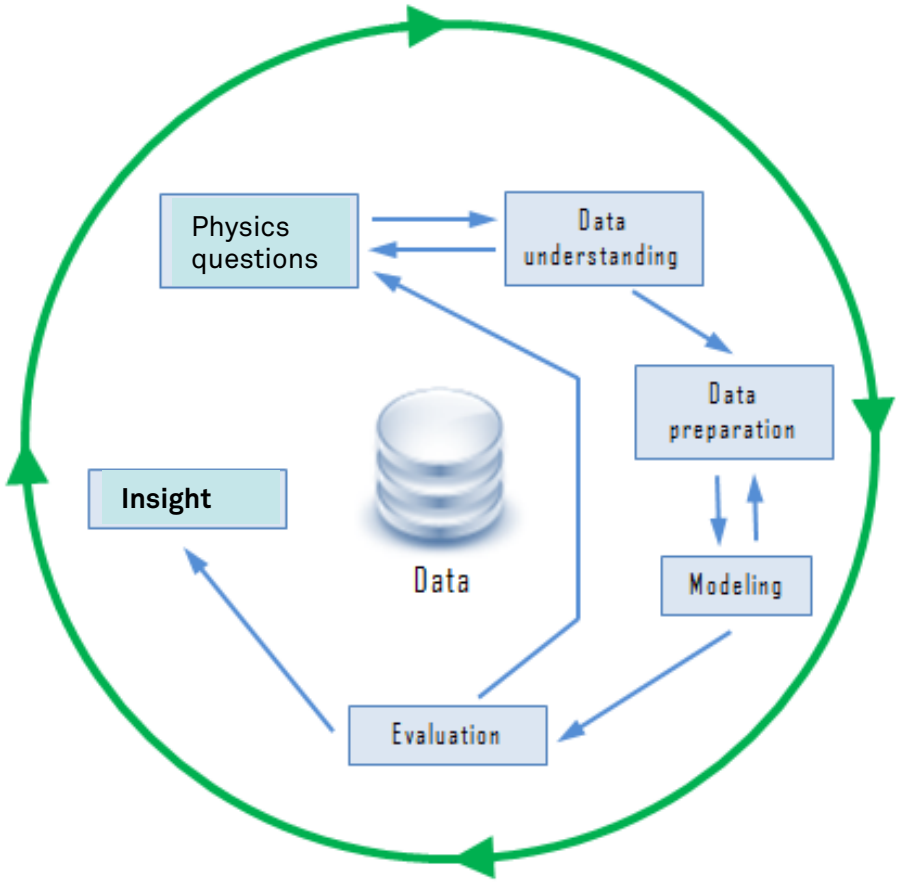
Overfitting to a sample? Measuring stability!

- Split the training data into m sets.
  - Do feature selection on m-1 sets.
  - Use the selected features for learning and test the learner's performance.
  - We do that m times.
- One sample leads to the feature set A, another to the feature set B, both select k features out of n given ones.
- Jaccard Index $J = \frac{|A \cap B|}{|A \cup B|}$

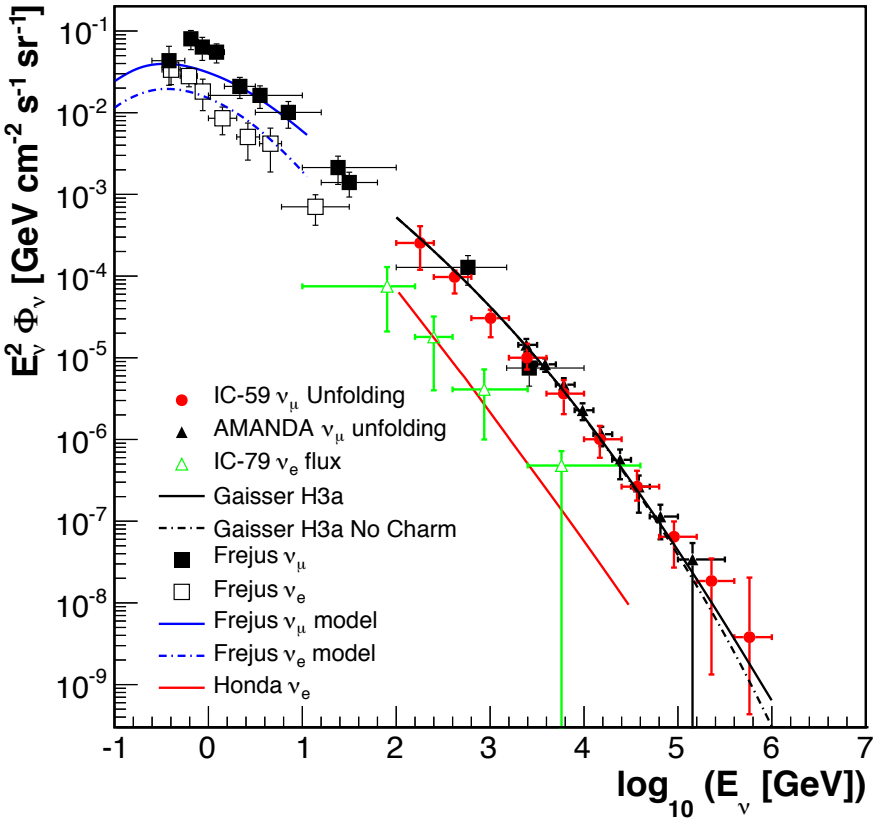


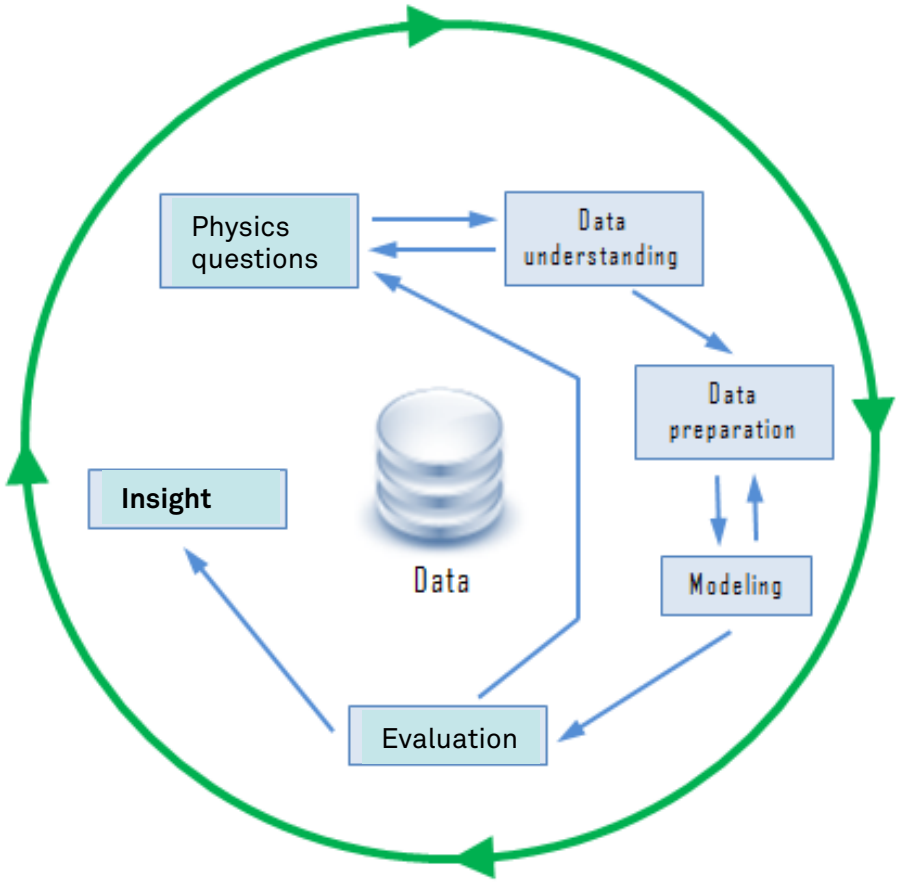
Offline Analysis with MRMR and Unfolding of IceCube Data

- Reconstructing the spectrum of atmospheric neutrinos could be up to an energy level 1 PeV (IceCube collaboration and Morik 2014).
- Quality cuts lead from full data D to D', rejecting the easy 91.4% of background.
- Random Forest leads from D' to D'' so that 99.9999% of background muons are rejected.
- At this background rejection 27 771 atmospheric neutrino events were detected in 346 days of IceCube 59.



Empirical work for theory development



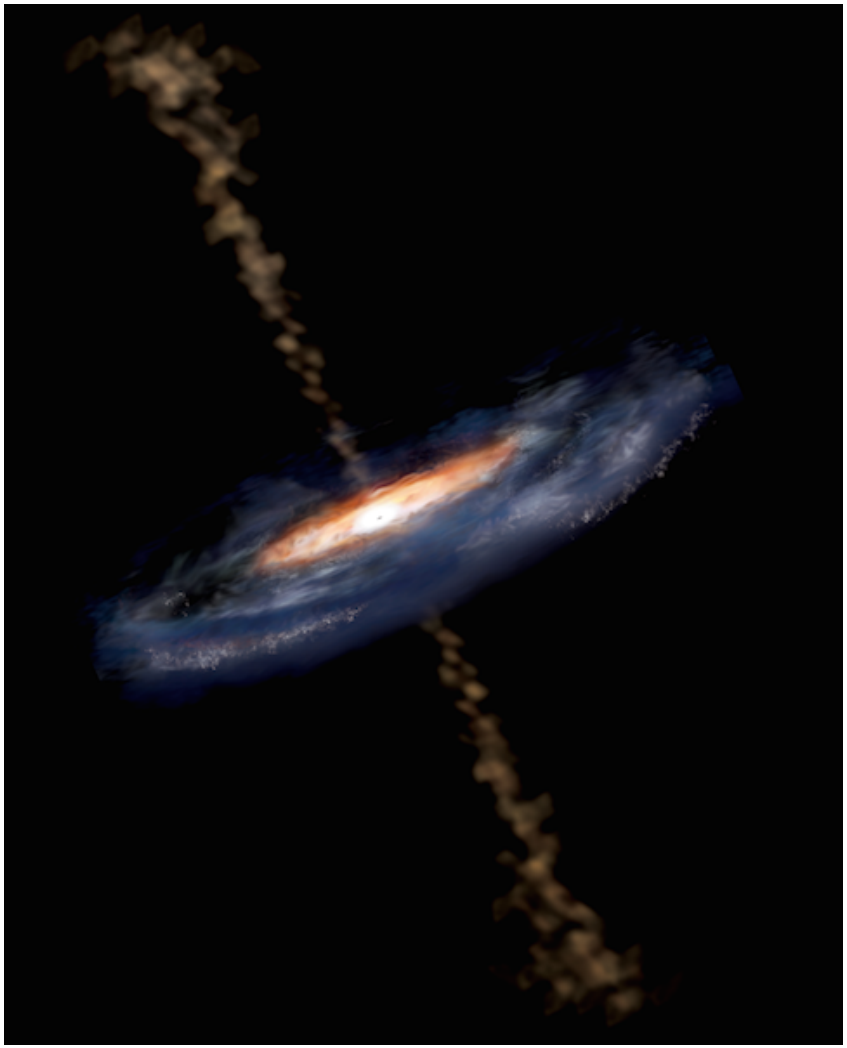


Overview

- Short introduction to the Collaborative Research Center SFB 876
- Tools for data analysis, streaming data
- Offline Data Analysis
  - IceCube
- Online Data Analysis
  - Magic, FACT

➢ Science today is based on data.

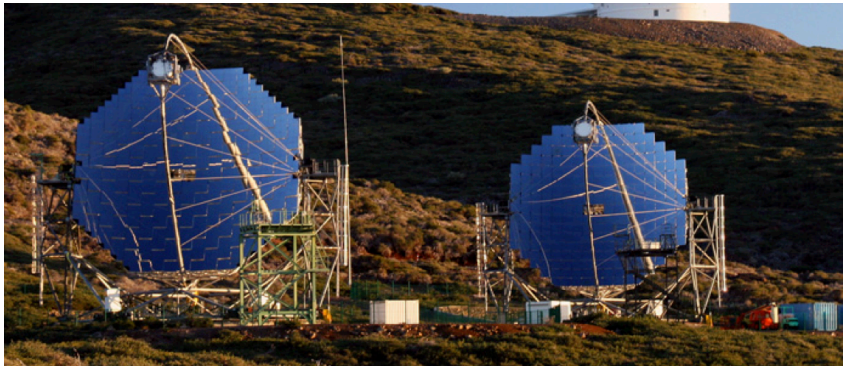
➢ Data analysis is intrinsically tied to the scientific process.



Active Galactic Nuclei

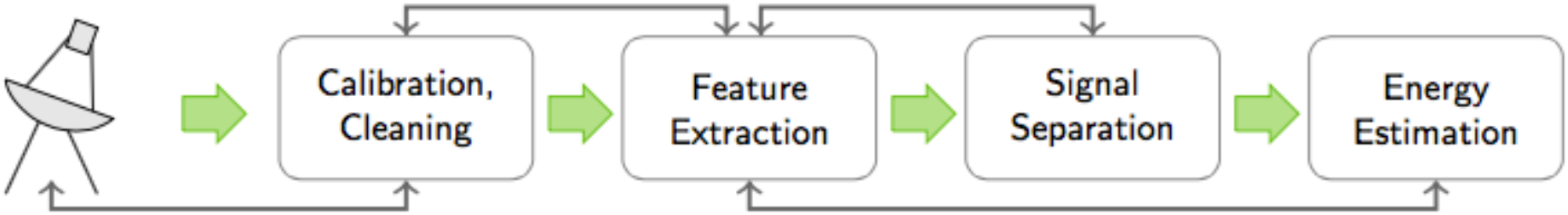
Skewed distribution challenging data analysis

- Calibration, cleaning
- Feature extraction
- Signal separation
- Energy estimation
- A simulator provides labeled observations.
- Gamma rays of high energy are rare events as opposed to hadrons, ratio 1 to 1000.



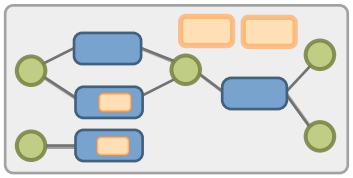
MAGIC I (2003) and MAGIC II (2009)
La Palma, Roque de los Muchachos
FACT telescope, same type, same place

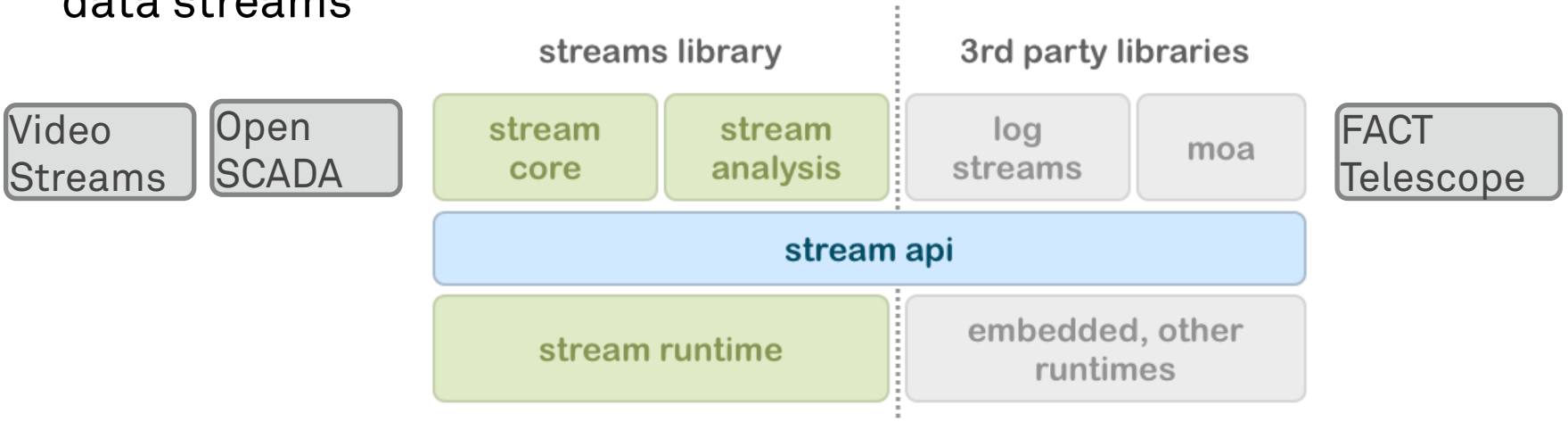
Bockermann, Christian and Brügge, Kai and Buss, Jens and Egorov, Alexey and Morik,Katharina and Rhode, Wolfgang and Ruhe, Tim 2015
"Online Analysis of High-Volume Data Streams in Astroparticle Physics"
Best Paper Award ECML PKDD 2015



Streams Framework for Real-Time Processing

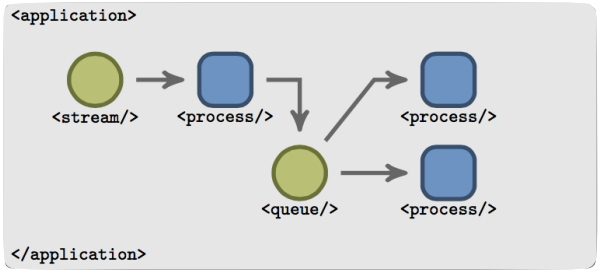
- Easing development
  - Use a simple XML specification for application data flows
  - Provide a simple API for custom code
- https://sfb876.de/streams/
- Middle layer framework for data streams

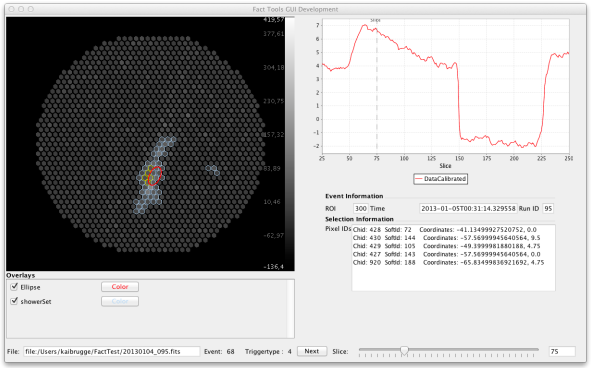




FACT Tools Library

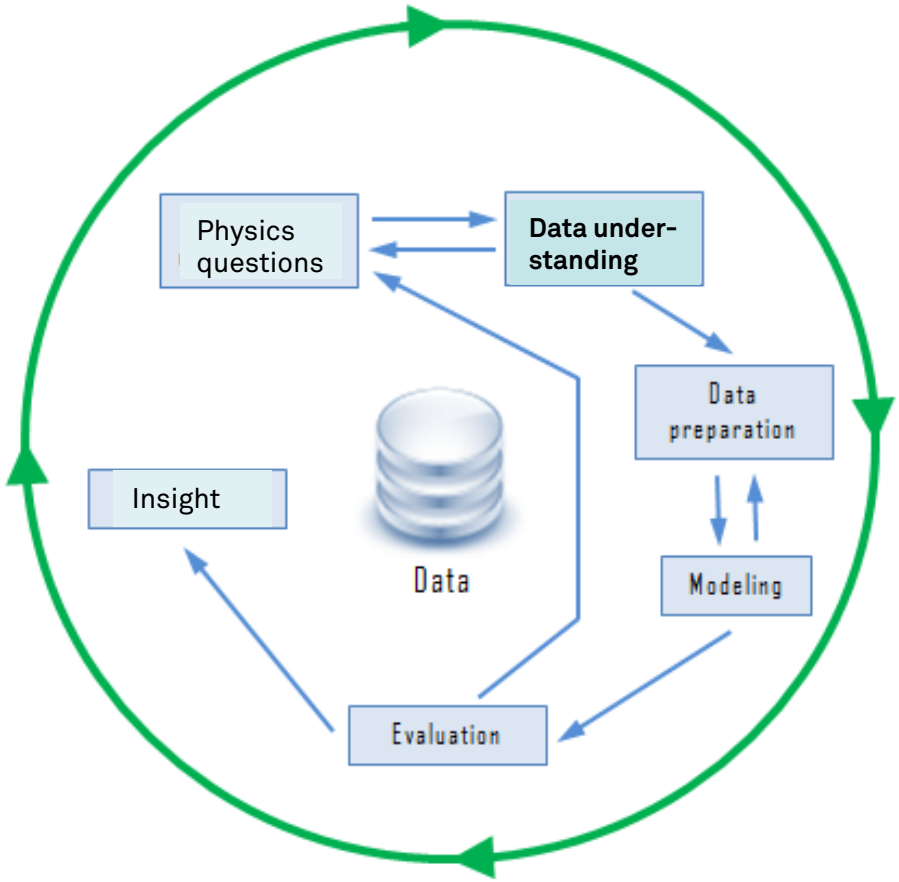
- Rapid prototyping of processing pipelines
- Testing new features, various classifiers
- Same pipeline for MC and data
- Declarative design in XML
  - Process sharing
  - Reproducibility of result
  - Integration of other tool boxes with CS theoretical validation



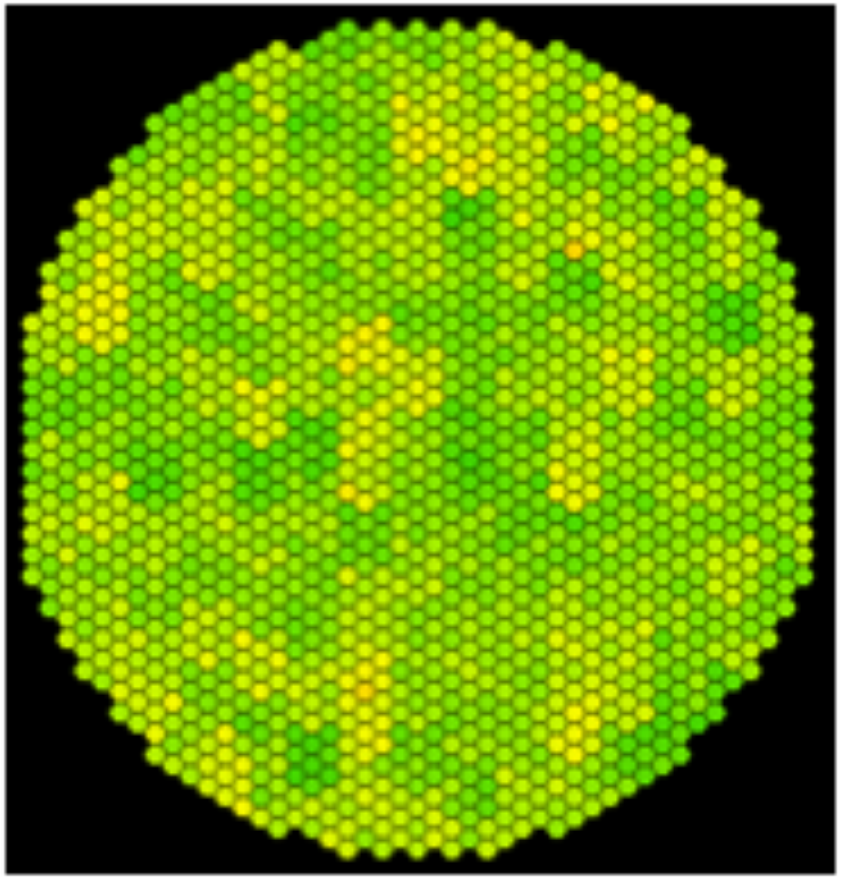


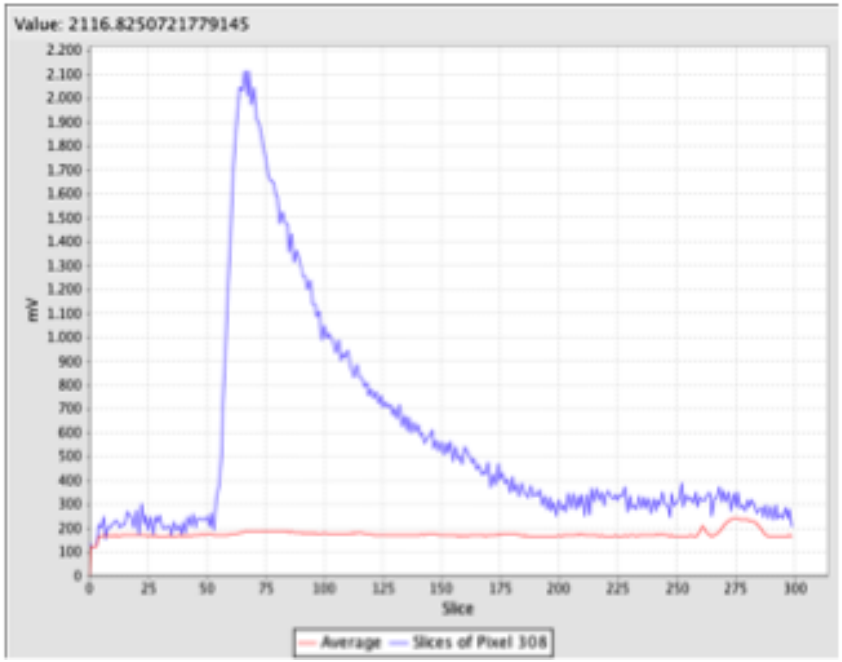
Data understanding

- What do the data look like?
- How to access the data?
- Do we need real-time access?

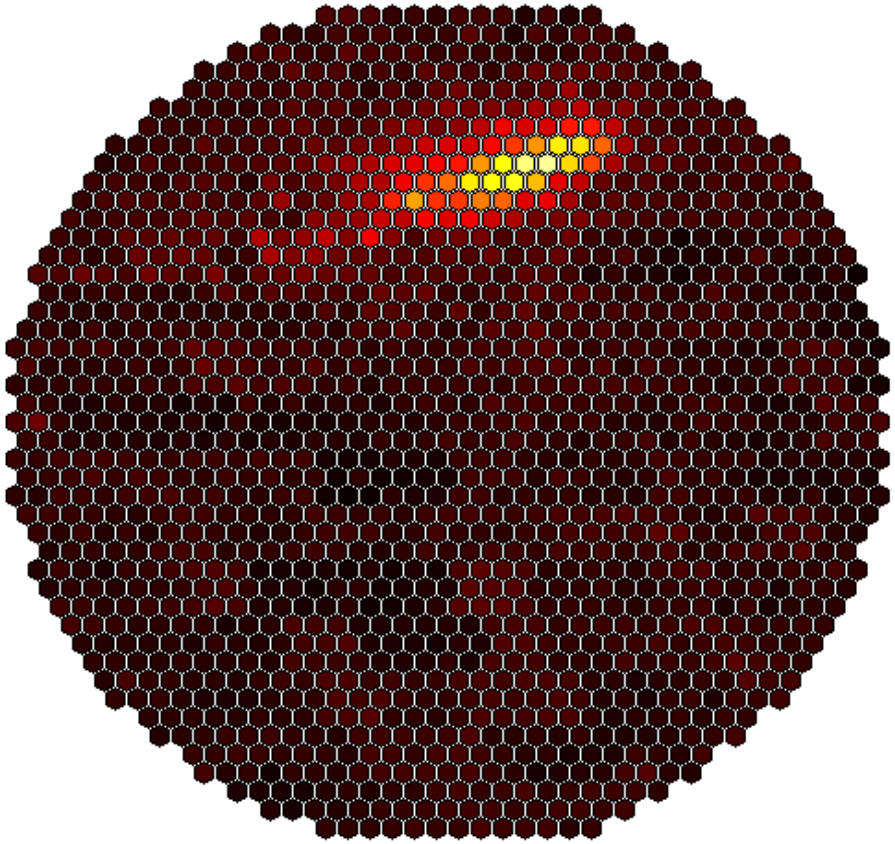


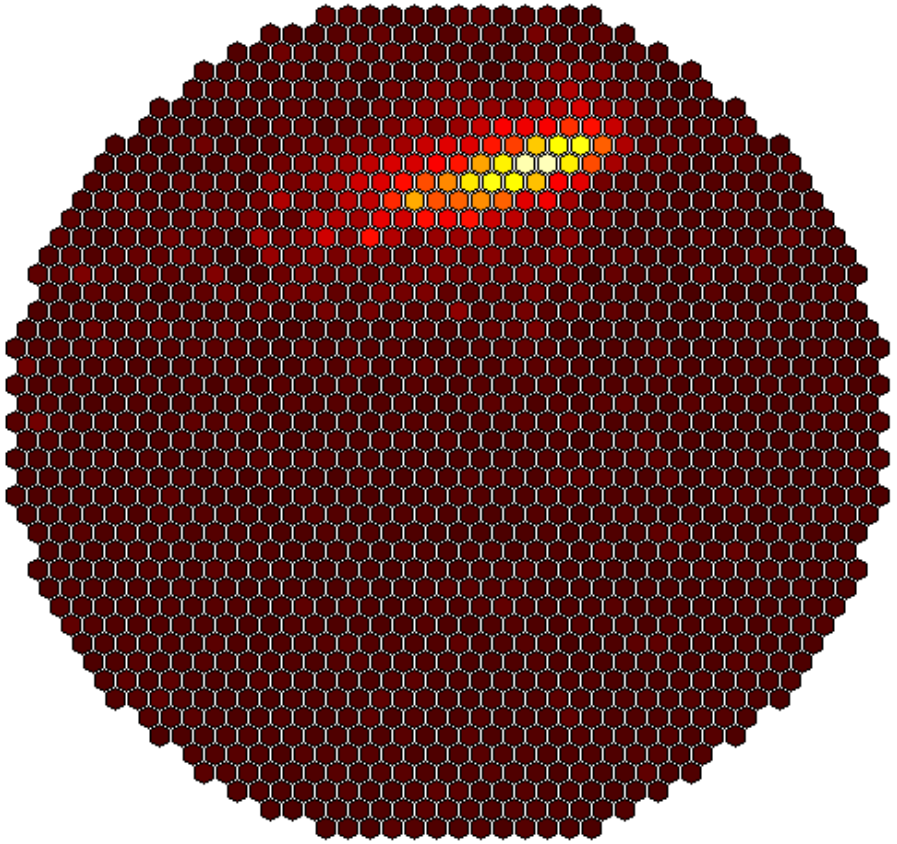
FACT-Viewer



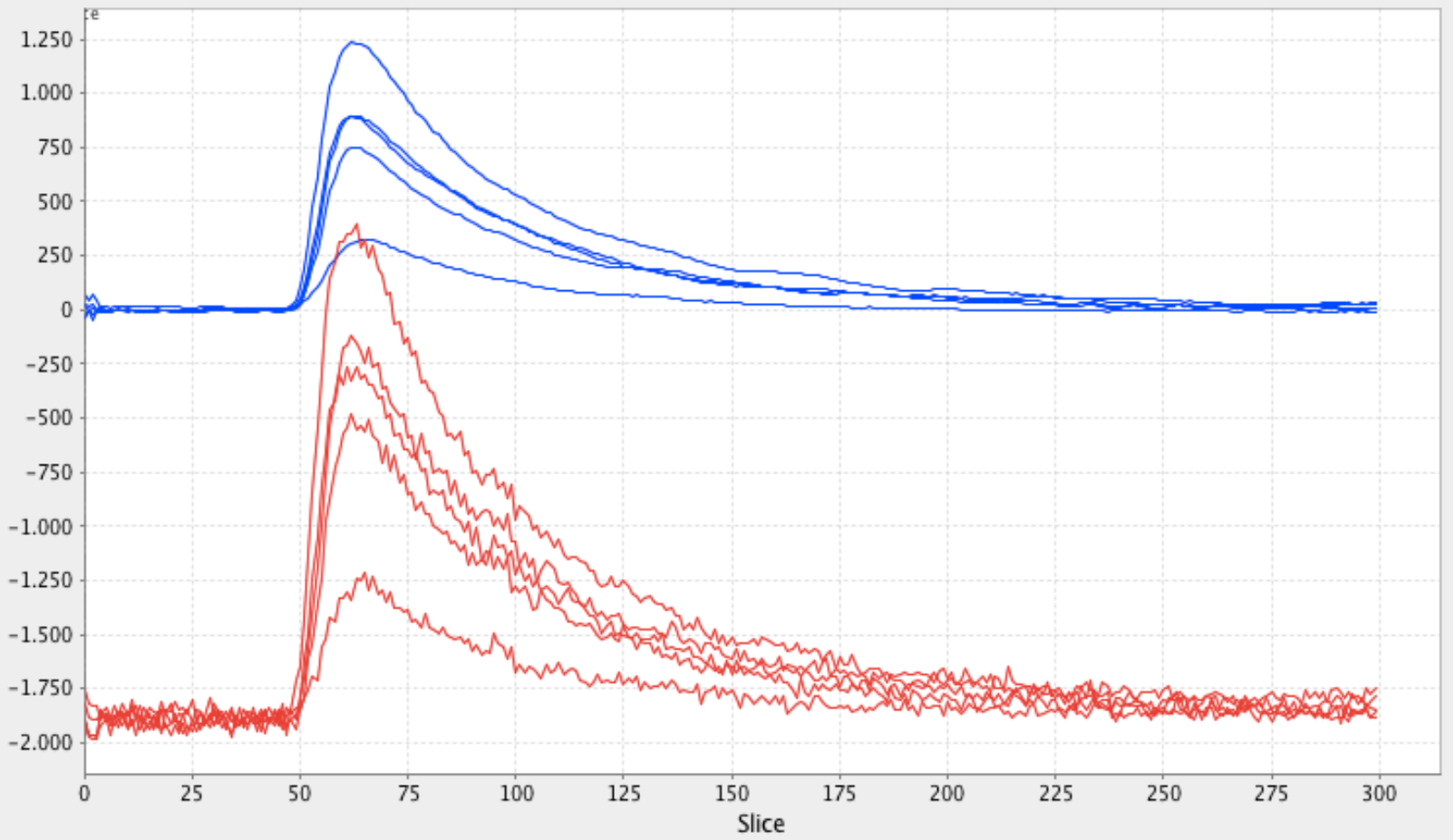

Value: 2116.8250721779145
mV
Slice
Average
Slices of Pixel 308

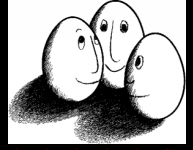
Data Calibration

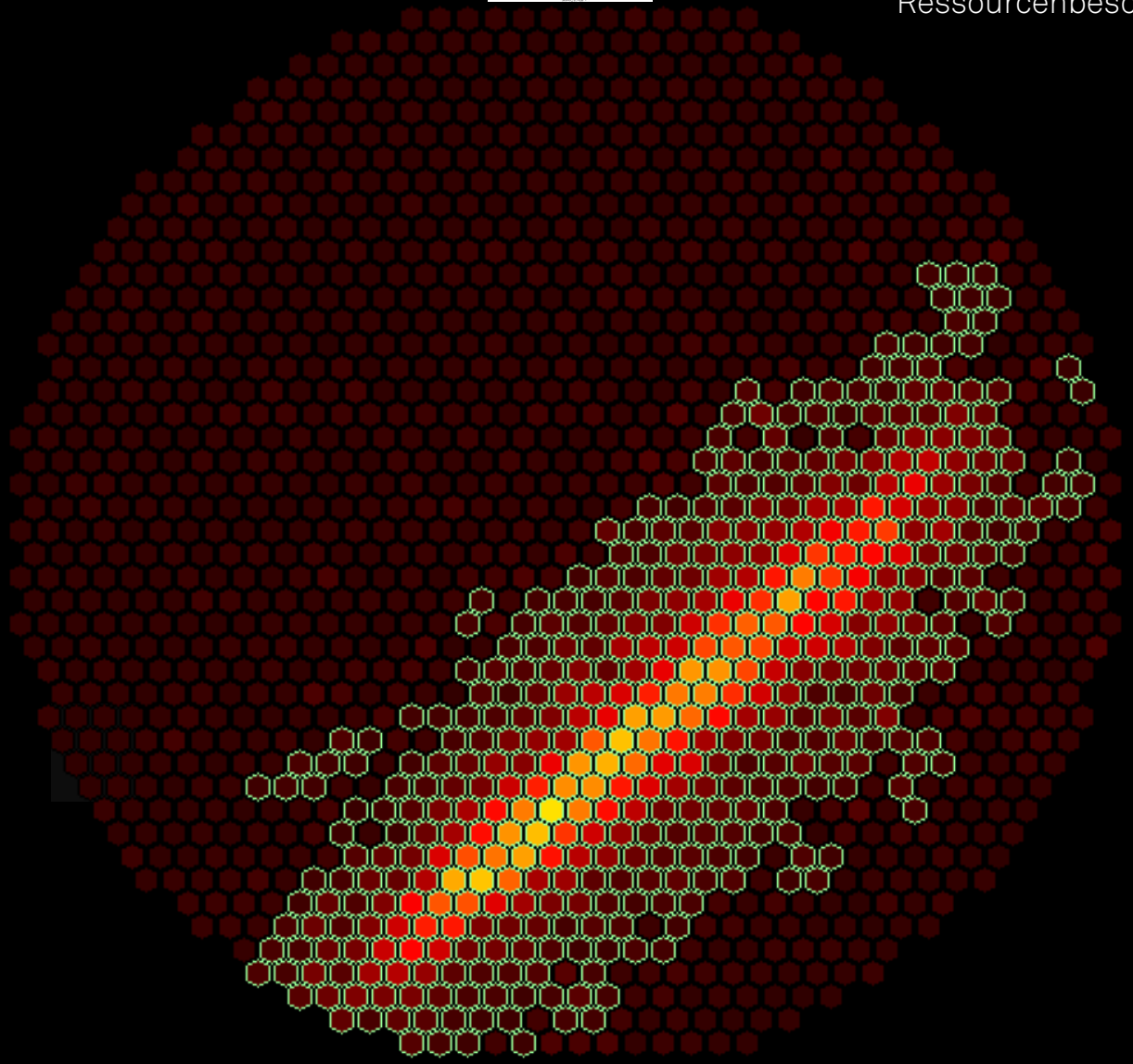




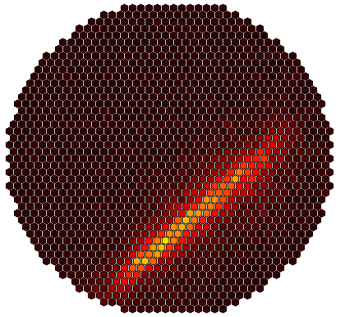
Data Calibration

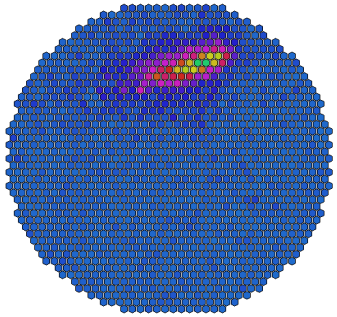




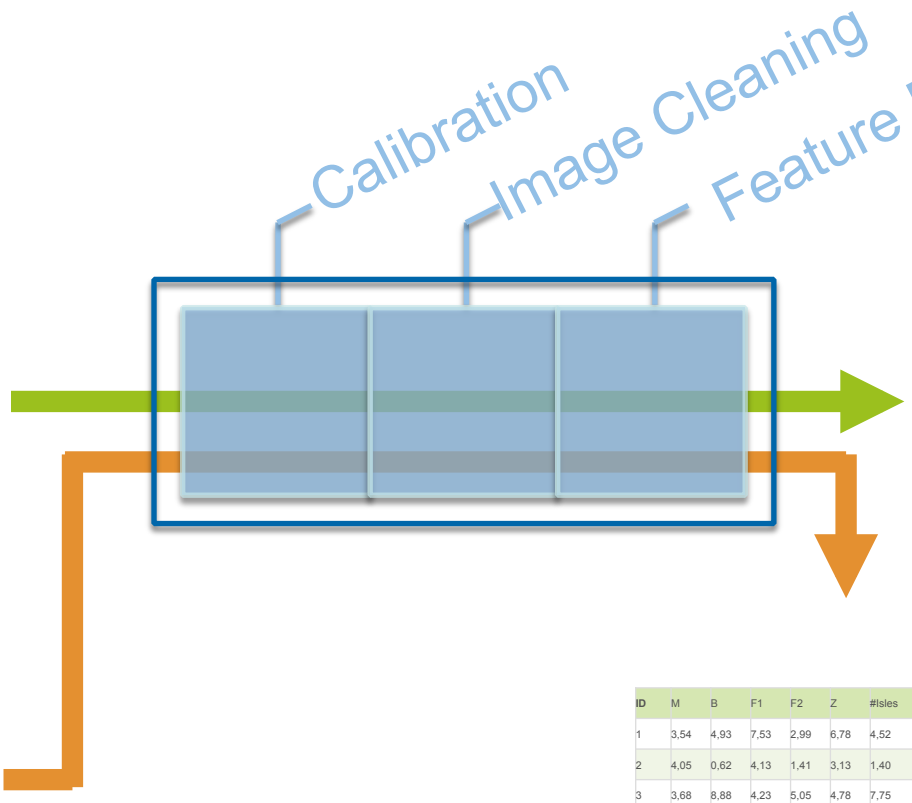


Raw Data





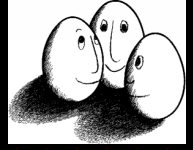
Simulated Raw Data

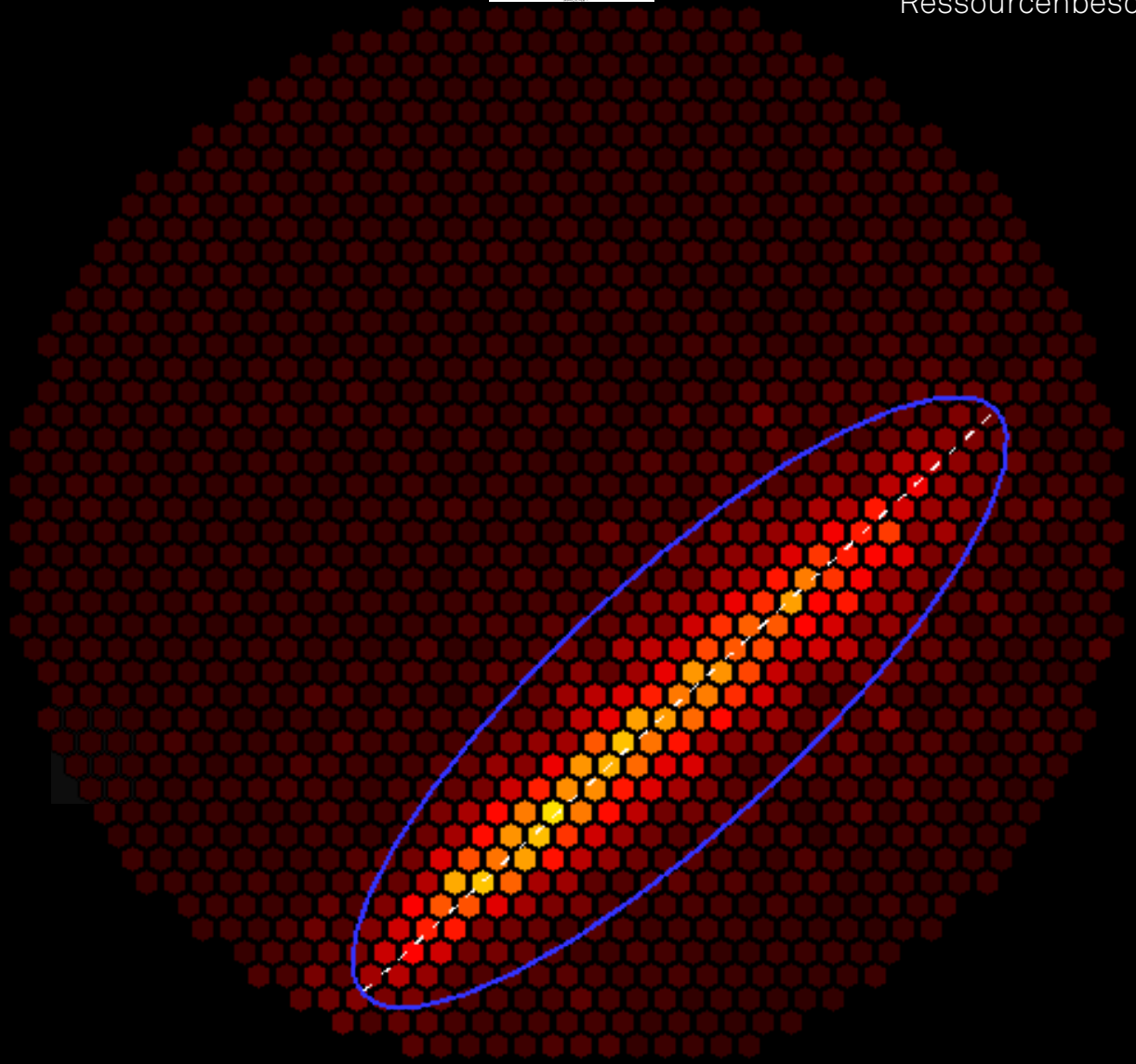


Feature Representation

ID	M	B	F1	F2	Z	#Isles	Width
1	3,54	4,93	7,53	2,99	6,78	4,52	5,64
2	4,05	0,62	4,13	1,41	3,13	1,40	1,98
3	3,68	8,88	4,23	5,05	4,78	7,75	8,48
4	6,75	4,58	0,48	3,79	6,37	3,81	1,76

ID	M	B	F1	F2	Z	#Isles	Width	Label
1	3,54	4,93	7,53	2,99	6,78	4,52	5,64	+1
2	4,05	0,62	4,13	1,41	3,13	1,40	1,98	-1
3	3,68	8,88	4,23	5,05	4,78	7,75	8,48	+1
4	6,75	4,58	0,48	3,79	6,37	3,81	1,76	-1

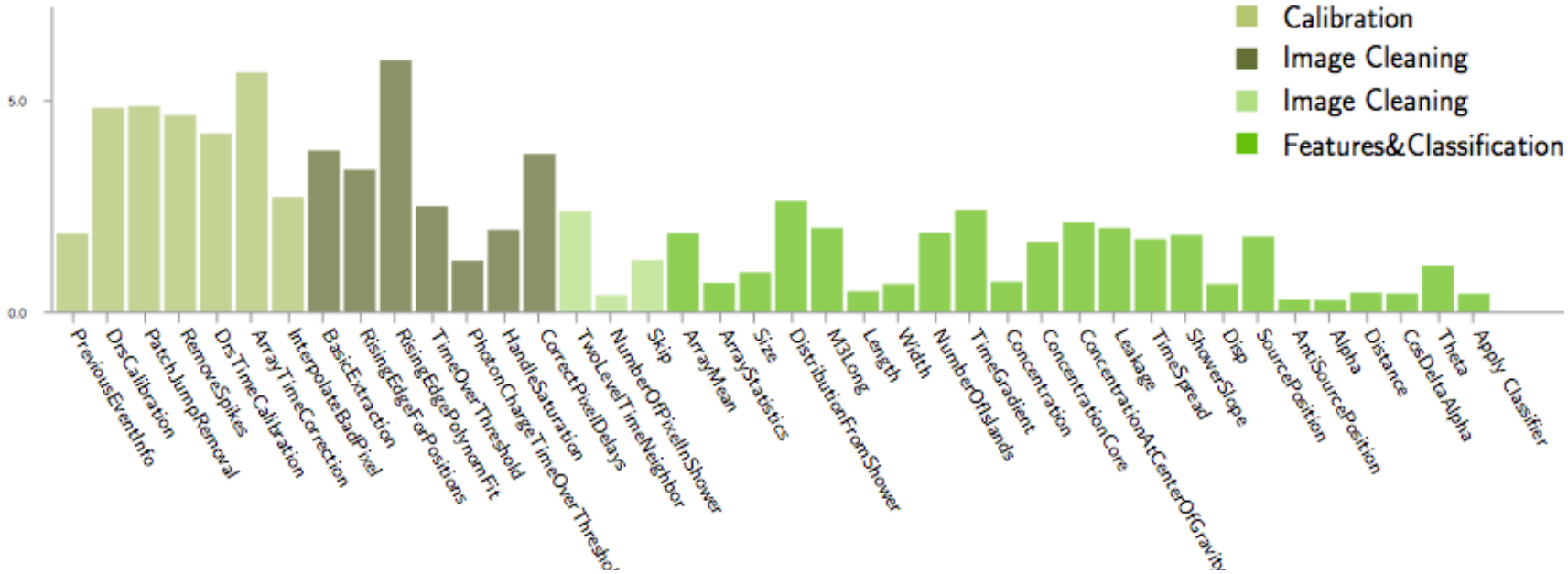




Width and **Length** of ellipse important features

Through-Put Performance of FACT Tools

- FACT records 60 events per second.
- Each events amounts to 3 Megabyte of raw data.
- ➢ 180MB/second are to be processed!
- ➢ Average processing time in milliseconds at a log scale shows the overall process ending with a classifier application.



CTA Project

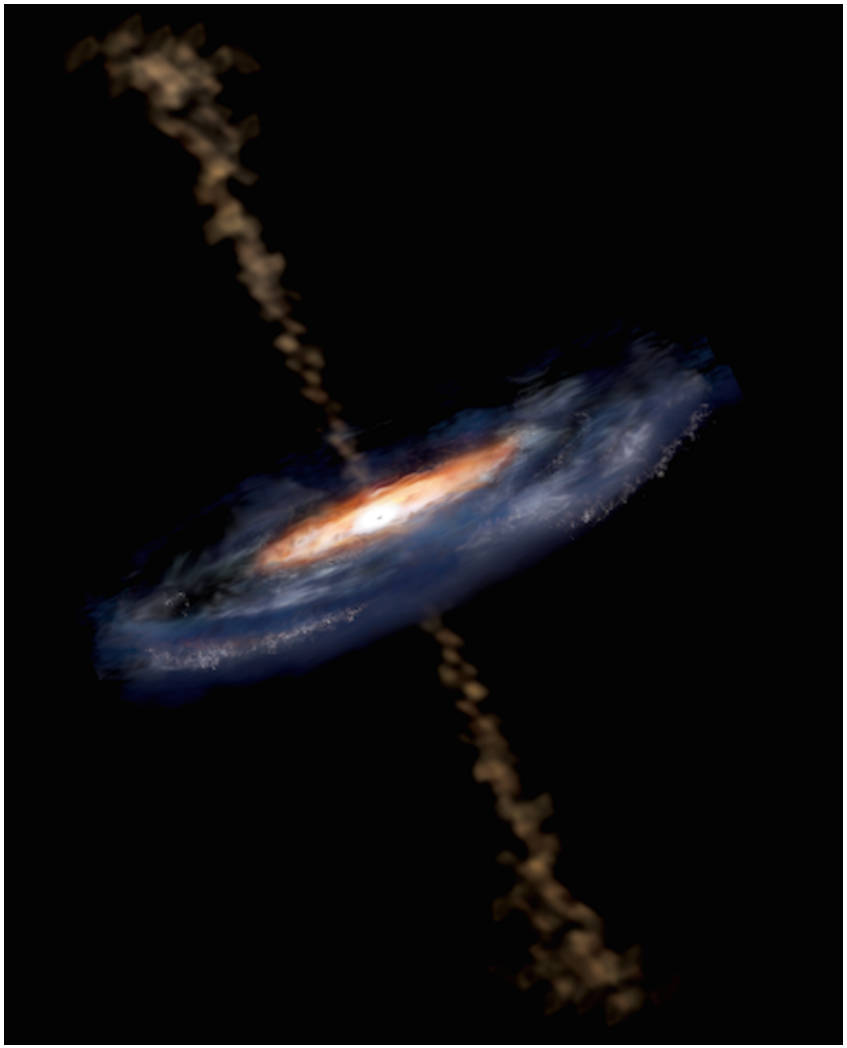
- CTA - Cherenkov Telescope Array
  - Two locations with lots of telescopes
    - North Site ~ 4 large, 15 medium telescopes
    - South Site ~ 4 large, 24 medium, 72 small telescopes
  - high sampling rates (> 10 kHz), high camera resolutions
  - combined analysis in high time resolution
  - domain experts design camera electronics, read-out boards, telescope construction, software (!?)

Overview

- Short introduction to the Collaborative Research Center SFB 876
- Tools for data analysis, streaming data
- Offline Data Analysis
  - IceCube
- Online Data Analysis
  - Magic, FACT

➢ Science today is based on data.

➢ Data analysis is intrinsically tied to the scientific process.

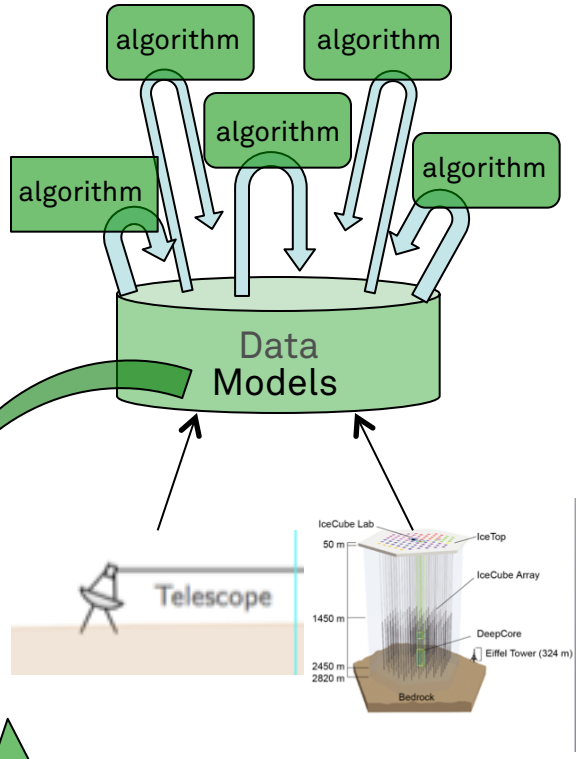


Active Galactic Nuclei

Data driven Science and Data Science

- Science is based on data, nowadays.
- Data are gathered in experiments.
- Data need to be cleaned, stored, summarized, and analyzed.
- The result of data processing should be a scientific insight.
- Each step needs a theoretical validation!
- Interdisciplinary:
  - Physics
  - Statistics
  - Computer science

Theory of Computer Science
Properties of algorithms,...

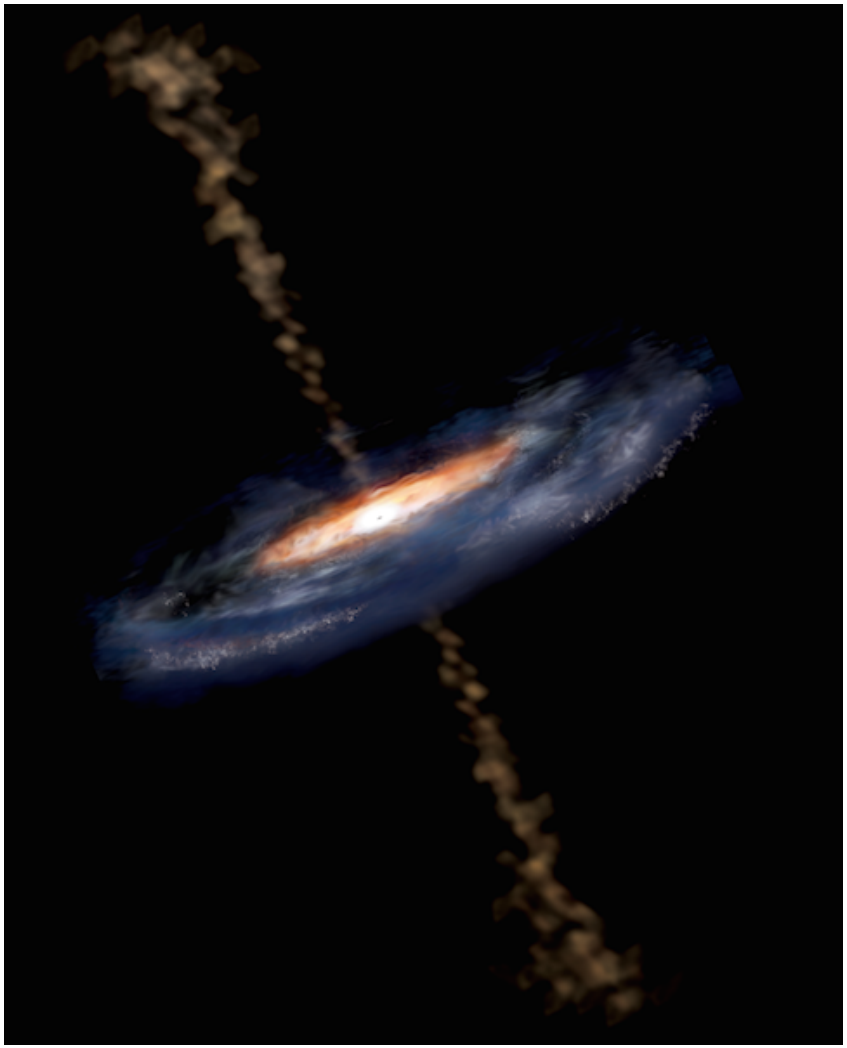


Theory of Statistics
Properties of vectors spaces, ...

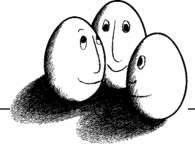
Theory of physics
Dark matter? Properties of matter...

Conclusion

- Data Analysis is needed in order to make good use of big data (in physics).
- We have seen for offline analysis RapidMiner.
- We have seen for online analysis streams framework.
- Choosing the right representation is the key to excellent results.
  - Stable MRMR feature selection(offline).
  - Data inspection and calibration (online).

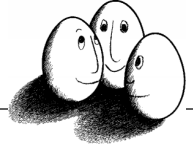


Active Galactic Nuclei



- **THANK YOU FOR YOUR ATTENTION!**

Data analysis is intrinsically tied to the scientific process.



Some References

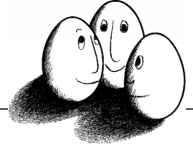
- Tim Ruhe, Katharina Morik, Benjamin Schowe (2011): Data Mining on Ice. In: Sarro, Bailer-Jones, Eyer, O'Mullane, de Ridder (eds), Procs. of the Workshop on Astrostatistics and Data Mining in Large Astronomical Databases.
- Tim Ruhe, Katharina Morik, Wolfgang Rhode (2011): Data Mining Ice Cubes. In: Gabriel et alii (eds), Astronomical Data Analysis Software and Systems ADASS.
- IceCube Collaboration und Katharina Morik (2014): Development of a General Analysis and Unfolding Scheme and its Application to Measure the Energy Spectrum of Atmospheric Neutrinos with IceCube. In: Eur. Phys.J. (2014).
- Christian Bockermann, Kai Brügge, Jens Buss, Alexey Agorov, Katharina Morik, Wolfgang Rhode, Tim Ruhe (2015) Online Analysis of High-Volume Data Streams in Astroparticle Physics. ECML PKDD Industrial Track, best industrial paper award
- Christian Bockermann “Mining Big Data Streams for Multiple Concepts”, Ph D thesis, TU Dortmund, 2015

References

- Mathis Börner, Wolfgang Rhode, Tim Ruhe, Katharina Morik (2015) Discovering Neutrinos through Data Analytics. In: ECML PKDD Nectar Track
- M.L.Ahnen, M. Balbo, M. Bergmann, A. Biland, C. Bockermann, T. Bretz, J. Buss, D. Dorner, S. Einecke , J. Freiwald , C. Hempfling , D. Hildebrand , G. Hughes , W. Lustermann, K. Mannheim , K. Meier , K. Morik, S. Müller , D. Neise , A. Neronov , M. Nöthe , A.K. Overkemping , A. Paravac , F. Pauss , W. Rhode , F. Temme , J. Thaele , S. Toscano , P. Vogler , R. Walter , A. Wilbert (2015) FACT-Tools: Streamed Real-Time Data Analysis.
  In: 34th Int. Cosmic Ray Conference, Procs of Science (PoS 865)
- M. Nöthe, M. L. Ahnen, M. Balbo, M. Bergmann, C. Bockermann, A. Biland, T. Bretz, K. A. Brügge, J. Buss, D. Dorner, S. Einecke, J. Freiwald, C. Hempfling, D. Hildebrand, G. Hughes, W. Lustermann, K. Mannheim, K. Meier, K. Morik, S. Müller, D. Neise, A. Neronov, A.-K. Overkemping, A. Paravac, F. Pauss, W. Rhode, F. Temme, J. Thaele, S. Toscano, P. Vogler, R. Walter, and A. Wilbert (2015) FACT – Calibration of Imaging Atmospheric Cherenkov Telescopes with Muon Rings.
  In: 34th Int. Cosmic Ray Conference, Procs of Science (PoS 733)

Some References

- F. Temme, M. L. Ahnen, M. Balbo, M. Bergmann, A. Biland, C. Bockermann, T. Bretz, K. A. Brügge, J. Buss, D. Dorner, S. Einecke, J. Freiwald, C. Hempfling, D. Hildebrand, G. Hughes, W. Lustermann, K. Mannheim, K. Meier, K. Morik, S. Müller, D. Neise, A. Neronov, M. Nöthe, A.-K. Overkemping, A. Paravac, F. Pauss, W. Rhode, J. Thaele, S. Toscano, P. Vogler, R. Walter, A. Wilbert (2015) FACT - First Energy Spectrum from a SiPM Cherenkov Telescope. In: 34th Int. Cosmic Ray Conference, Procs of Science (PoS 707)
- Buss, Jens and Ahnen, M.L. and Balbo, M. and Bergmann, M. and Biland, Adrian and Bockermann, Christian and Bretz, Thomas and Brügge, K. A. and Dorner, Daniela and Einecke, Sabrina and Freiwald, Jan and Hempfling, Christina and Hildebrand, D. and Hughes, Gareth and Lustermann, Werner and Mannheim, Karl and Meier, K. and Morik, Katharina and Müller, Sebastian and Neise, Dominik and Neronov, A. and Nöthe, Max and Rhode, Wolfgang and Temme, Fabian and Thaele, Julia and Toscano, Simona and Vogler, Patrick and Walter, Roland and Wilbert, A. (2015): FACT - Influence of SiPM Crosstalk on the Performance of an Operating Cherenkov Telescope



Some References

- Jens Buß, Christian Bockermann, Katharina Morik (2016) FACT-Tools – Processing High-Volume Telescope Data. In: Astronomical Data Analysis Software and Systems Conference, Trieste

A quote illustrating that developing algorithms for statistical equations is a challenge:

- Signal inference with unknown response: Calibration-uncertainty renormalized estimator by Sebastian Dorn1,2,∗, Torsten A. Enßlin1,2, Maksim Greiner1,2, Marco Selig1,2, and Vanessa Boehm1,2
  1 Max-Planck-Institut für Astrophysik Garching, Germany
  2 Ludwigs-Maximilians-Universität München, Germany
  (arxiv: March 3, 2015)

„Since this work is supposed to be a proof of concept we work with explicit matrices and tensors, whereby we have to limit the size of the problem for computational reasons. Further investigations are needed on how to transform this into a method using implicit tensors, and therefore suitable for “big data” problems. „