Hadoop in Social Network Analysis

- overview on tools and some best practices -

GridKa School 2013, Karlsruhe | 2013-08-27

Mirko Kämpf | mirko@cloudera.com

MARTIN-LUTHER UNIVERSITÄT HALLE-WITTENBERG

SOCIONICAL

cloudera
Ask Bigger Questions

HADOOP DEVELOPMENT TOOLS

Mirko Kämpf

Physicist, TU Chemnitz, 2009
Java Trainer, since 2003
Java Developer, since 1996
Committer, PPMC @ ASF

Hadoop Trainer, Cloudera, Inc.

Research Project:
SOCIONICAL, Martin-Luther Universität Halle-Wittenberg
Open Source Activity:
Hadoop Development Tools (Apache HDT)
Hadoop.TS (on GITHUB)

WHATS COMMING?

1) Complex Systems, from Time Series to Networks ...

2) Data, data, and even more data ... but how to handle it?

3) Some results of our project ...

4) Lessons learned, some recommendations ...

images from Google image search ...

Hadoop in Social Network Analysis

Abstract:

A Hadoop cluster is the tool of choice for many large scale analytics applications. A large variety of commercial tools is available for typical SQL like or data warehouse applications, but how to deal with **networks** and **time series**?

How to **collect and store data** for social media analysis and what are good practices for working with libraries like Mahout and Giraph?

The sample use case deals with a data set from Wikipedia to **illustrate** how to combine multiple public data sources with personal data collections, e.g. from Twitter or even personal mailboxes. We discuss efficient approaches for **data organisation**, data **preprocessing** and for **time dependent** graph analysis.

Bulk Processing

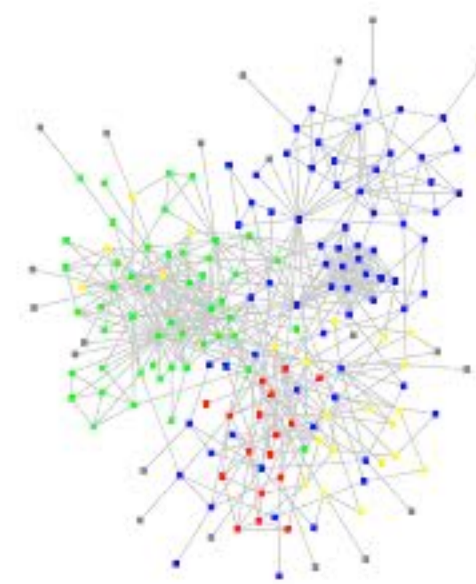
Datatypes Pig **I/O-Formats**

SequenceFile **OpenTSDB** Hive **HBase**

Distributed Storage Random-Access

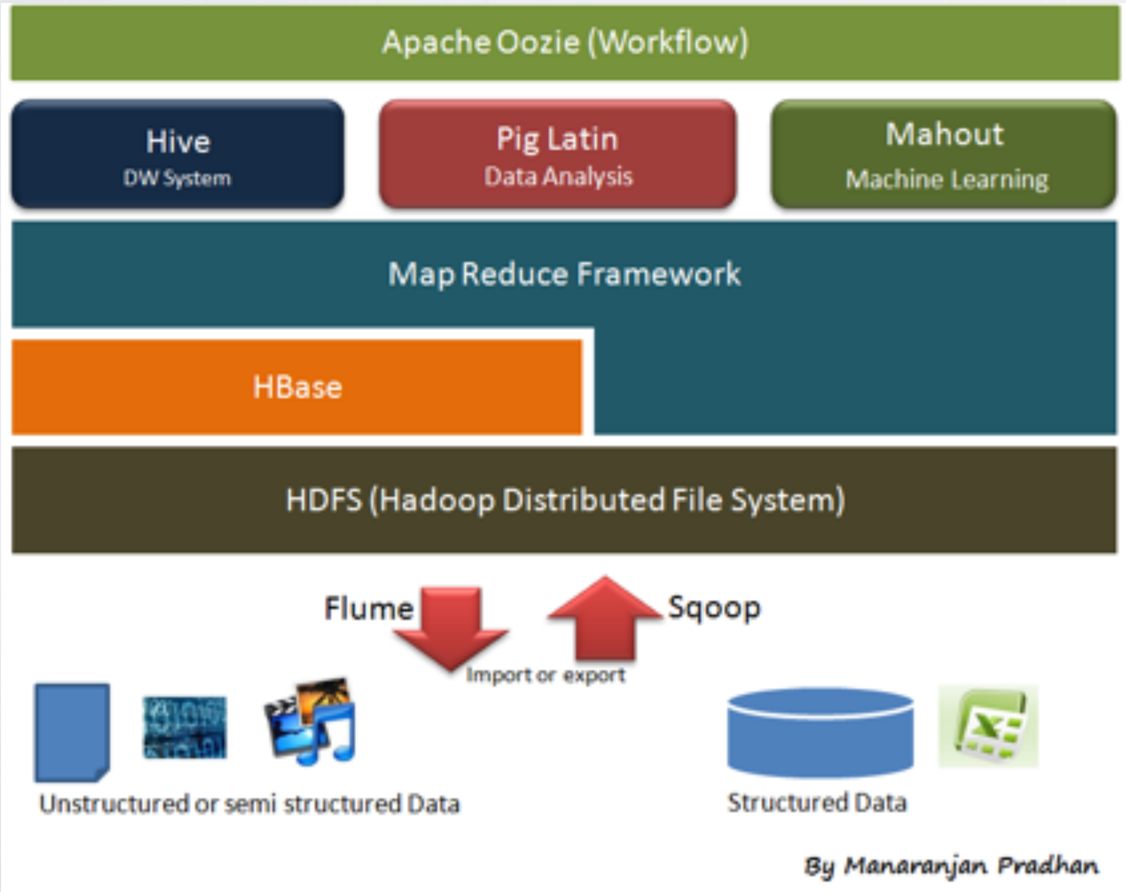
Sampling Rate VectorWriteable

Partitions



Hadoop in Social Network Analysis

Abstract:
A Hadoop cluster is the tool of choice for many large scale analytics applications. A large variety of commercial tools is available for typical SQL like or data warehouse applications, but how to deal with **networks** and **time series**?
How to **collect and store data** for social media analysis and what are good practices for working with libraries like Mahout and Giraph?
The sample use case deals with a data set from Wikipedia to **illustrate** how to combine multiple public data sources with personal data collections, e.g. from Twitter or even personal mailboxes. We discuss efficient approaches for **data organisation**, data **preprocessing** and for **time dependent** graph analysis.



(A) The Hadoop Ecosystem, offers a new technology to store and process large data sets, which are in the focus of interdisciplinary research.

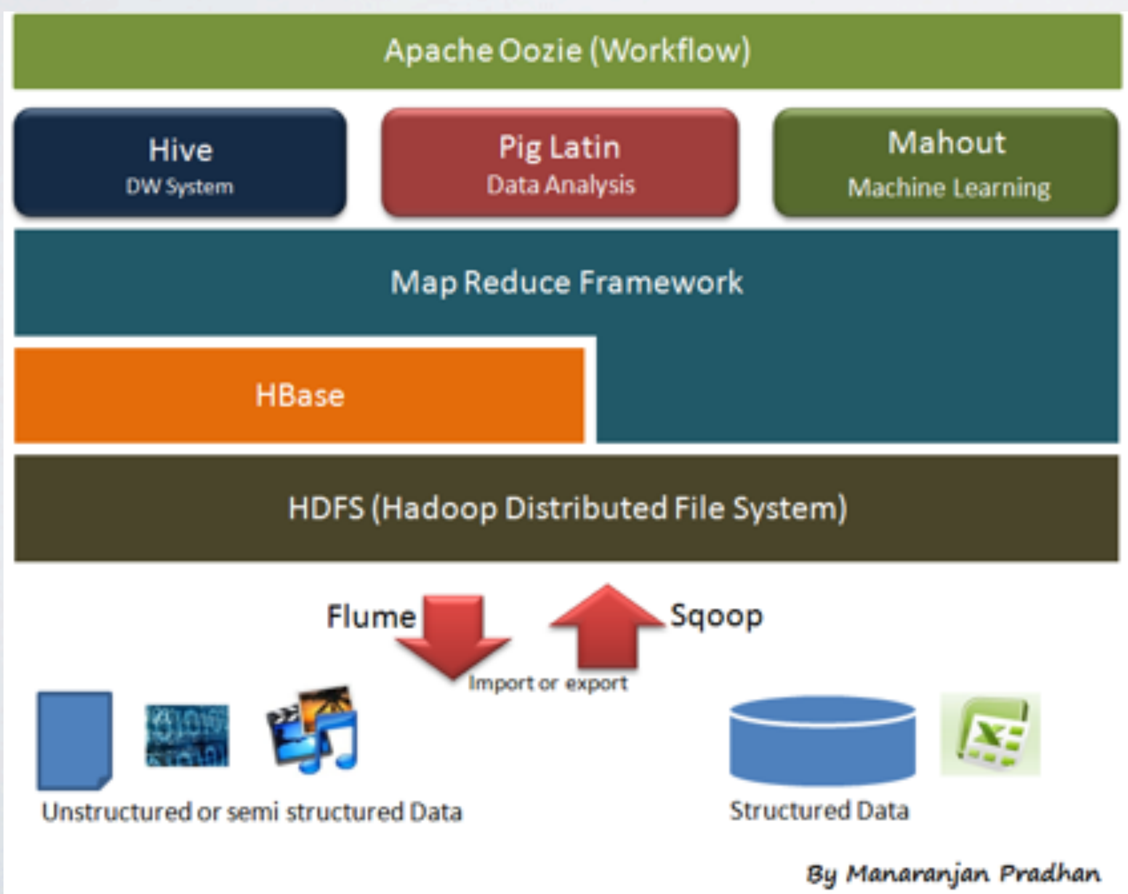
(B) Our data sets are created or generated by highly dynamic and flexible Social Media Applications.

(C) This requires new scientific approaches from complex systems research and also new technology.

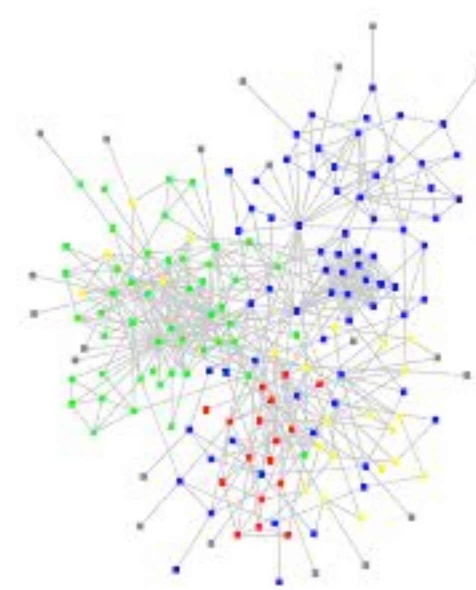
... and the loop is cosed.

Hadoop in Social Network Analysis

Abstract:
A Hadoop cluster is the tool of choice for many large scale analytics applications. A large variety of commercial tools is available for typical SQL like or data warehouse applications, but how to deal with **networks** and **time series**?
How to **collect and store data** for social media analysis and what are good practices for working with libraries like Mahout and Giraph?
The sample use case deals with a data set from Wikipedia to **illustrate** how to combine multiple public data sources with personal data collections, e.g. from Twitter or even personal mailboxes. We discuss efficient approaches for **data organisation**, data **preprocessing** and for **time dependent** graph analysis.



Social networks consist of <u>nodes</u>, which are the real world **objects** and <u>edges</u>, which are e.g. **relations**, **interactions** or **dependencies** between nodes.



Complex Networks

Definitions:

"A system comprised of a (usually large) number of (usually strongly) **interacting** entities, processes, or agents, ...

the understanding of which requires the development, or the use of, **new scientific tools**, nonlinear models, out-of equilibrium descriptions and computer simulations." [Advances in Complex Systems Journal]

"A system that can be analyzed into many components having relatively many relations among them, so **that the behavior of each component depends** on the behavior of others." [Herbert Simon]

"A system that involves numerous interacting agents whose **aggregate behaviors** are to be understood. Such aggregate activity is nonlinear, hence it **cannot simply be derived from summation** of individual components behavior." [Jerome Singer]

Nonlinear models

out-of equilibrium

Aggregation

Dynamics of Components

Dynamics of Subsystems

Interaction

Hierarchical Systems

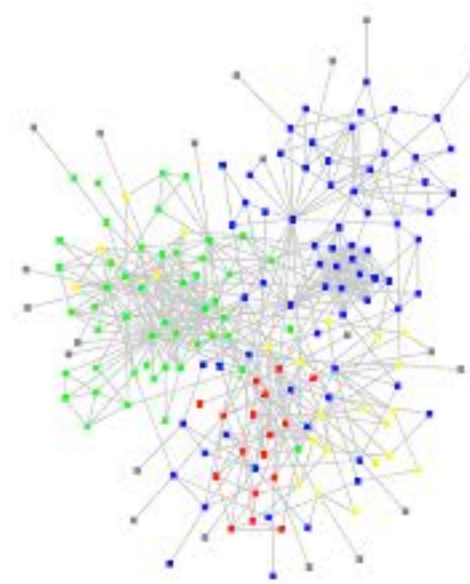
Superposition not possible

Dependency cycles

Based on Rocha, Luis M. [1999]. *BITS: Computer and Communications News*.
Computing, Information, and Communications Division. Los Alamos National Laboratory. Nov. 1999.

images from Google image search ...

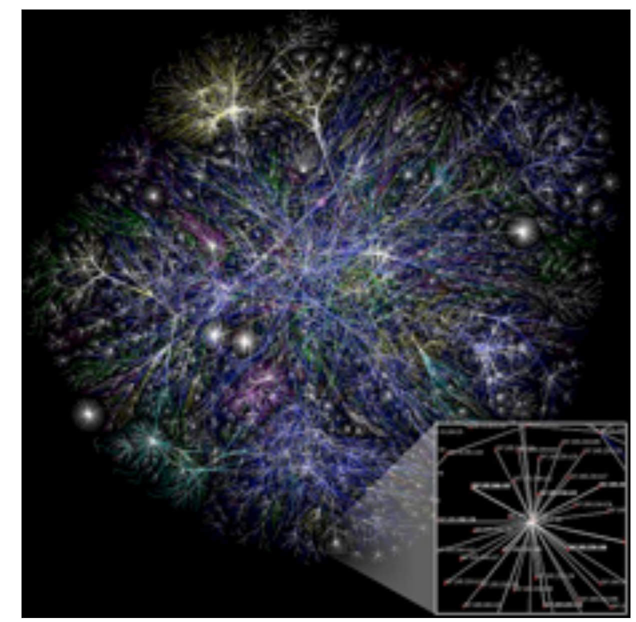
Social Networks are Complex Networks.



INTRODUCTION OF OUR PROJECT

Social online-systems are complex systems used for, e.g., information spread.

We develop and apply tools from time series analysis and network analysis to study the static and dynamic properties of social on-line systems and their relations.

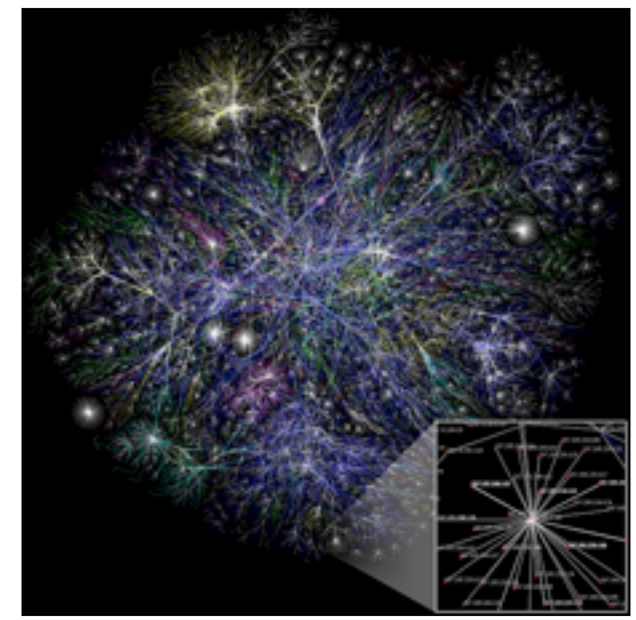


INTRODUCTION OF OUR PROJECT

Social online-systems are complex systems used for, e.g., information spread.

We develop and apply tools from time series analysis and network analysis to study the static and dynamic properties of social on-line systems and their relations.

Webpages (the nodes of the WWW) are linked in different, but related ways:



direct links pointing from one page to another (binary, directional)
similar access activity (cross-correlated time series of download rates)
similar edit activity (synchronized events of edits or changes)

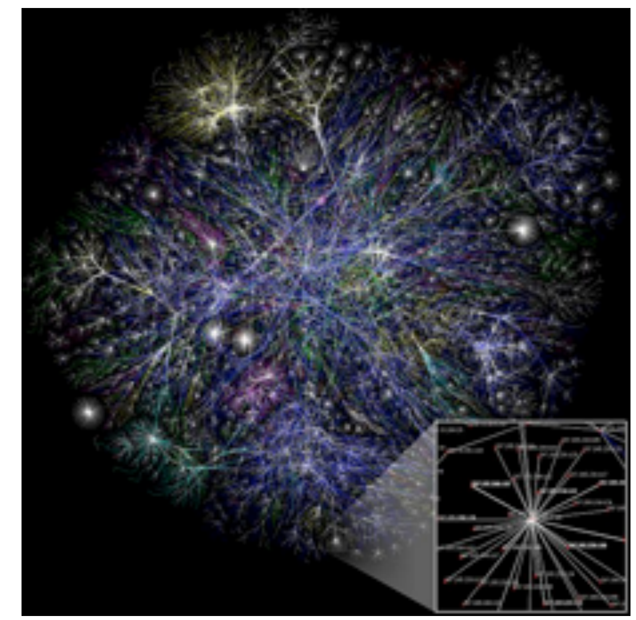
We extract the time-evolution of these three networks from real data. Nodes are identical for all three studied networks, but links and network structure as well as dynamics are different. We quantify how the inter-relations and inter-dependencies between the three networks change in time and affect each other.

INTRODUCTION OF OUR PROJECT

Social online-systems are complex systems used for, e.g., information spread.

We develop and apply tools from time series analysis and network analysis to study the static and dynamic properties of social on-line systems and their relations.

Webpages (the nodes of the WWW) are linked in different, but related ways:



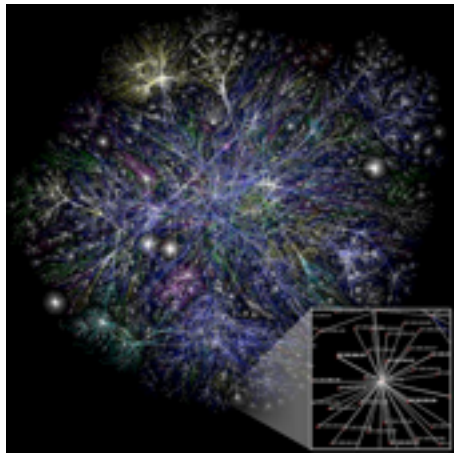
direct links pointing from one page to another (binary, directional)
similar access activity (cross-correlated time series of download rates)
similar edit activity (synchronized events of edits or changes)

We extract the time-evolution of these three networks from real data. Nodes are identical for all three studied networks, but links and network structure as well as dynamics are different. We quantify how the inter-relations and inter-dependencies between the three networks change in time and affect each other.

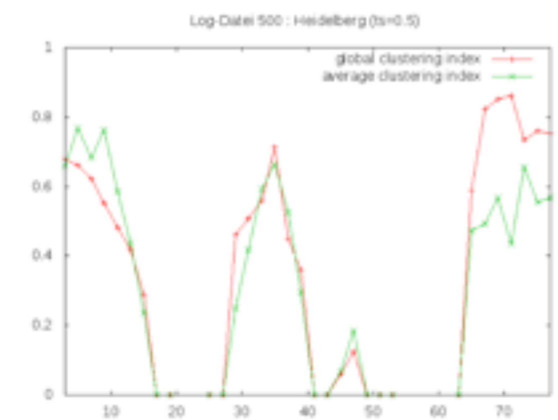
Example: Wikipedia → **reconstruct co-evolving networks**

1. Cross-link network between articles (pages, nodes)
2. Access behavior network (characterizing similarity in article reading behavior)
3. Edit activity network (characterizing similarities in edit activity for each article)

CHALLENGES ...



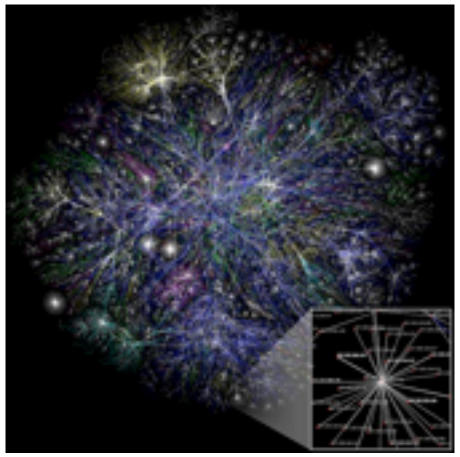
Complex System

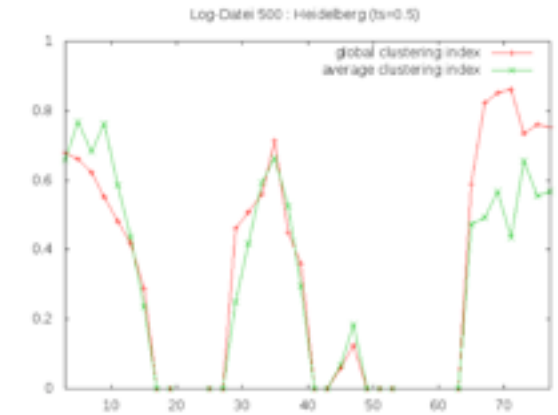


Time evolution ???

- Data points (time series) collected at independent locations or obtained form individual objects do not show dependencies directly.
- It is a common task, to calculate several types of correlations, but how are these results affected by special properties of the raw data?
- What meaning do different correlations have and how can we eliminate artifacts of the calculation method?

CHALLENGES ...



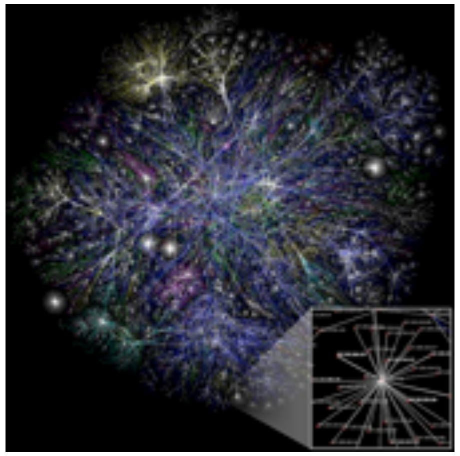


Complex System

Time evolution ???

- Data points (time series) collected at independent locations or obtained form individual objects do not show dependencies directly.
- It is a common task, to calculate several types of correlations, but how are these results affected by special properties of the raw data?
- What meaning do different correlations have and how can we eliminate artifacts of the calculation method?

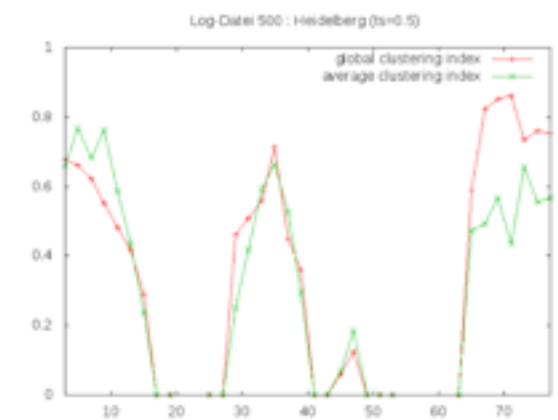
CHALLENGES ...



Complex System

Element properties

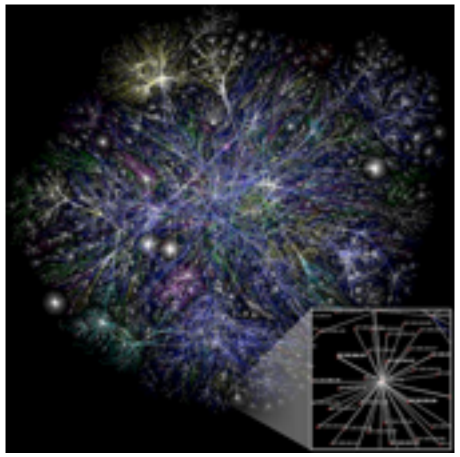
Measured data is "disconnected"



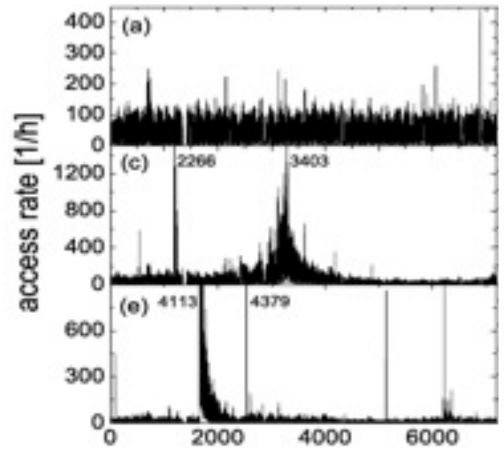
Time evolution ???

- Data points (time series) collected at independent locations or obtained form individual objects do not show dependencies directly.
- It is a common task, to calculate several types of correlations, but how are these results affected by special properties of the raw data?
- What meaning do different correlations have and how can we eliminate artifacts of the calculation method?

CHALLENGES ...

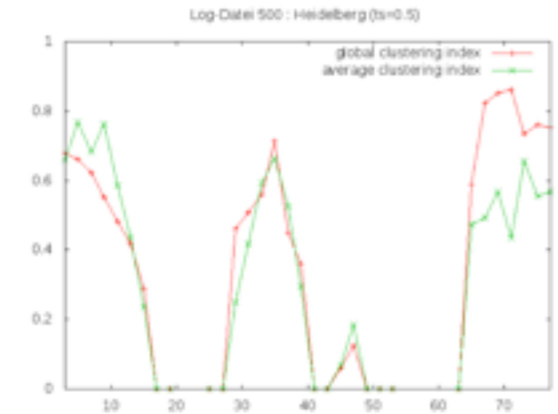


Complex System



Element properties

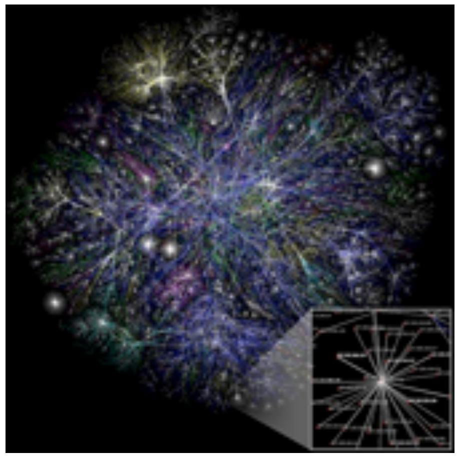
Measured data is "disconnected"



Time evolution ???

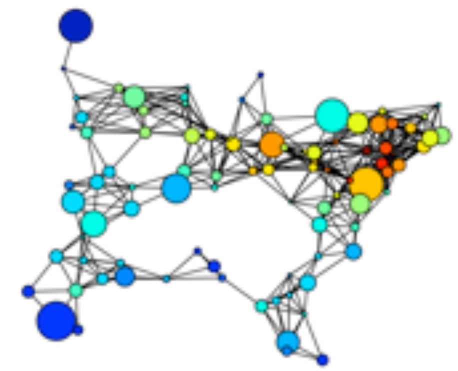
- Data points (time series) collected at independent locations or obtained form individual objects do not show dependencies directly.
- It is a common task, to calculate several types of correlations, but how are these results affected by special properties of the raw data?
- What meaning do different correlations have and how can we eliminate artifacts of the calculation method?

CHALLENGES ...

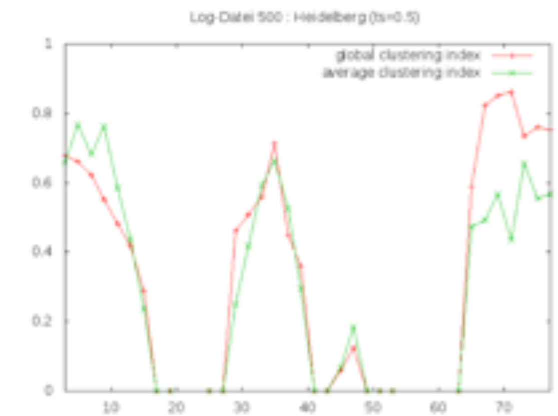


Complex System

Element properties
Measured data is "disconnected"



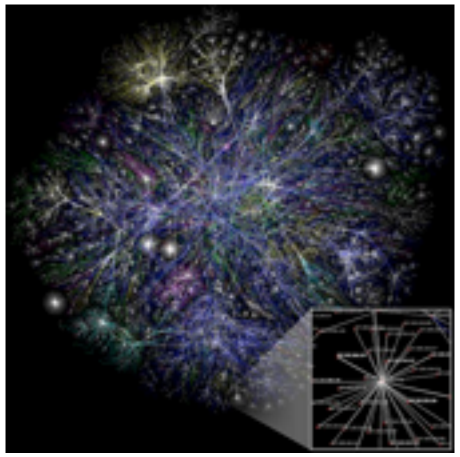
System properties
Derived from relations between elements and structure of the network



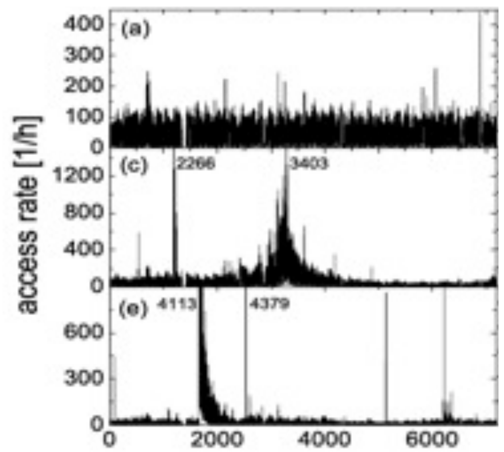
Time evolution ???

- Data points (time series) collected at independent locations or obtained form individual objects do not show dependencies directly.
- It is a common task, to calculate several types of correlations, but how are these results affected by special properties of the raw data?
- What meaning do different correlations have and how can we eliminate artifacts of the calculation method?

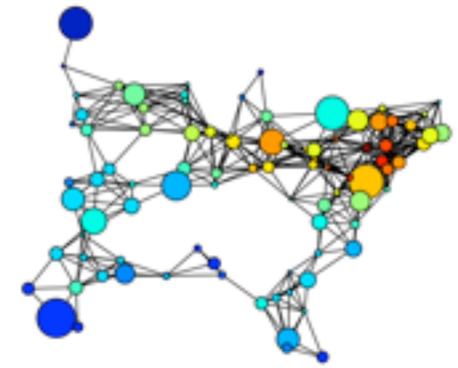
CHALLENGES ...



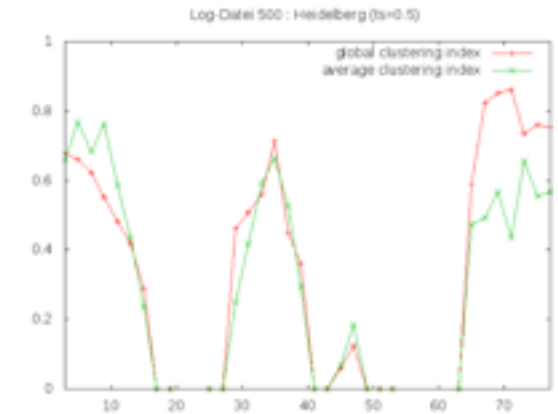
Complex System



Element properties
Measured data is "disconnected"

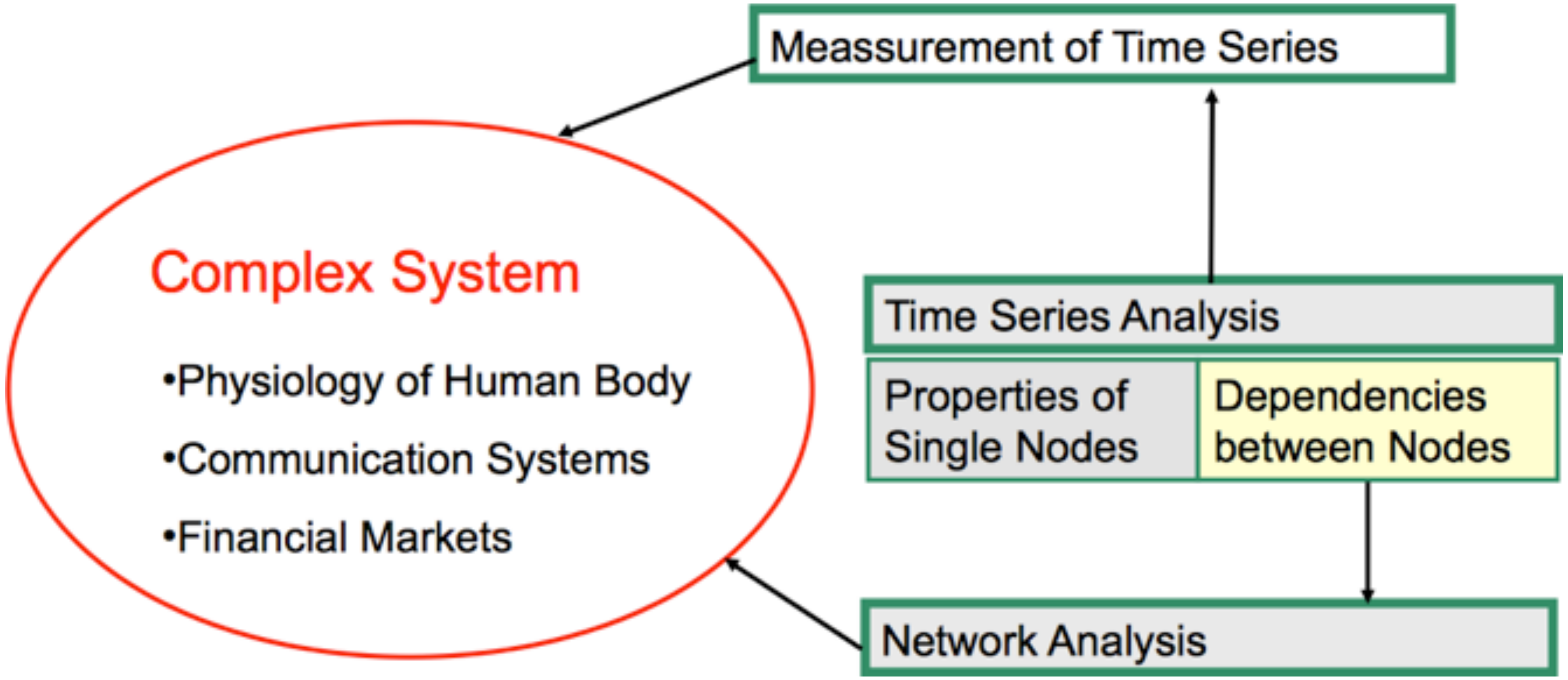


System properties
Derived from relations between elements and structure of the network



Time evolution ???

- Data points (time series) collected at independent locations or obtained form individual objects do not show dependencies directly.
- It is a common task, to calculate several types of correlations, but how are these results affected by special properties of the raw data?
- What meaning do different correlations have and how can we eliminate artifacts of the calculation method?



- Time Series Analysis
  - if our data set is well prepared and we have records with **well defined properties** (as in RDBMS), than Hive and Pig work well.
  - How to organize the loose data in records?
  - How to deal with sliding windows?
  - How to handle intermediate data?

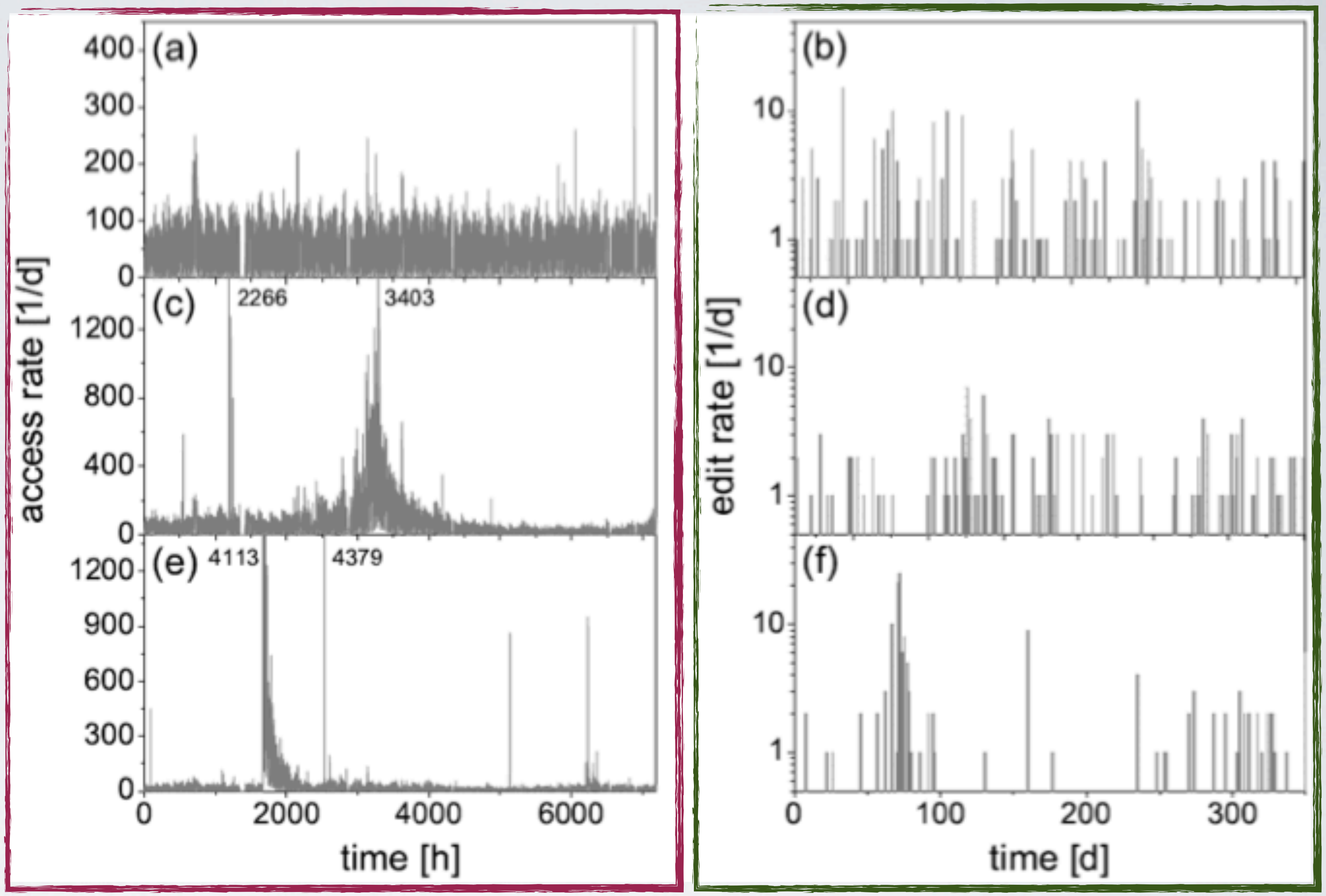
TIME SERIES: WIKIPEDIA USER ACTIVITY

Node = article
(specific topic)

Available data:

1. **Hourly access frequency**
(number of article downloads for each hour in ≈ 300 days)

2. **Edit events**
(time stamps for all changes in the wikipedia pages)



Examples of Wikipedia access time series for three articles with (a,b) stationary access rates ('Illuminati (book)'), (c,d) an endogenous burst of activity ('Heidelberg'), and (e,f) an exogenous burst of activity ('Amoklauf Erfurt'). The left parts show the complete hourly access rate time series (from January 1, 2009, till October 21, 2009; i.e. for 42 weeks = 294 days = 7056 hours). The right parts show edit-event data for the three representative articles.

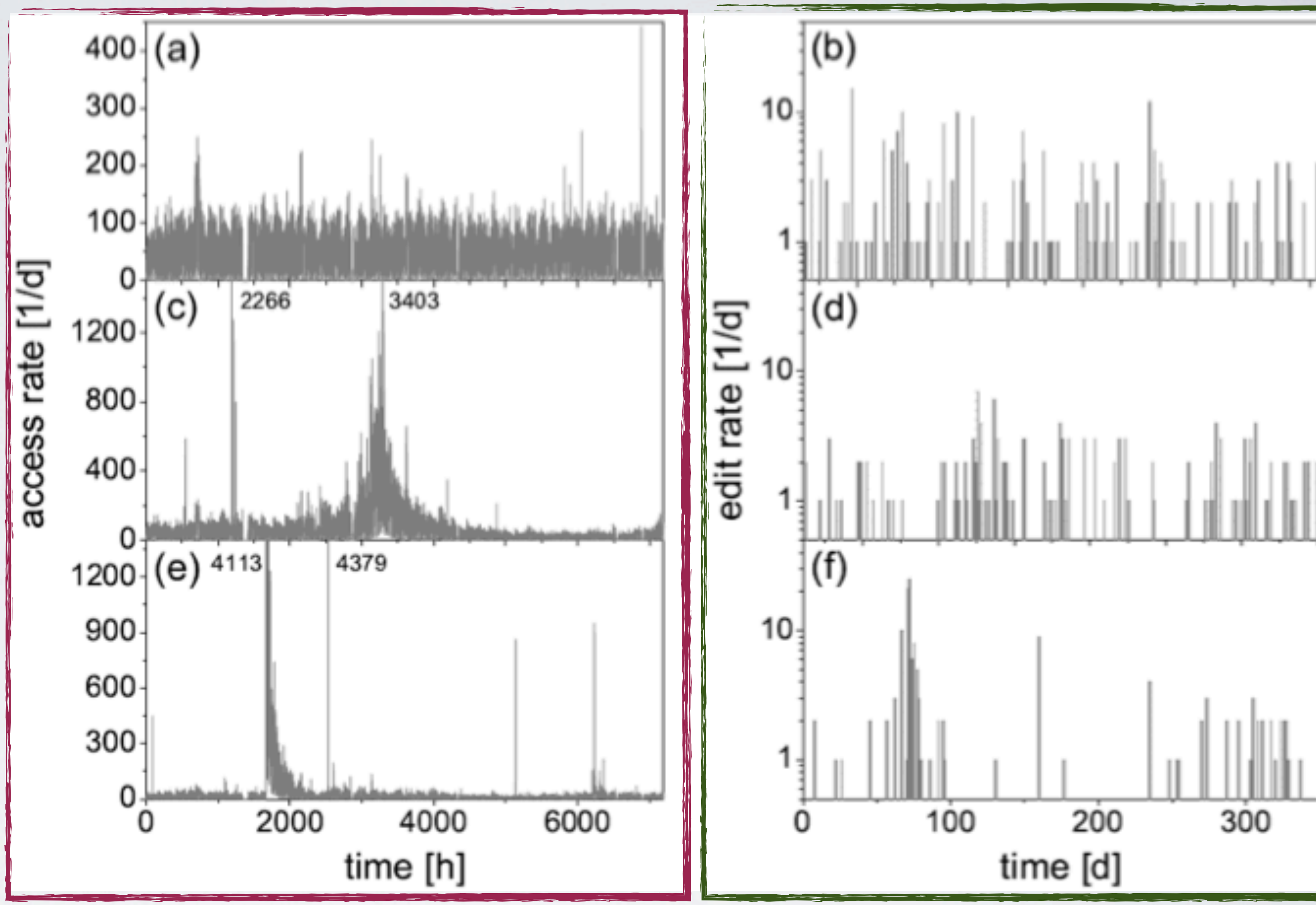
TIME SERIES: WIKIPEDIA USER ACTIVITY

Node = article
(specific topic)

Available data:

1. **Hourly access frequency**
(number of article downloads for each hour in ≈ 300 days)

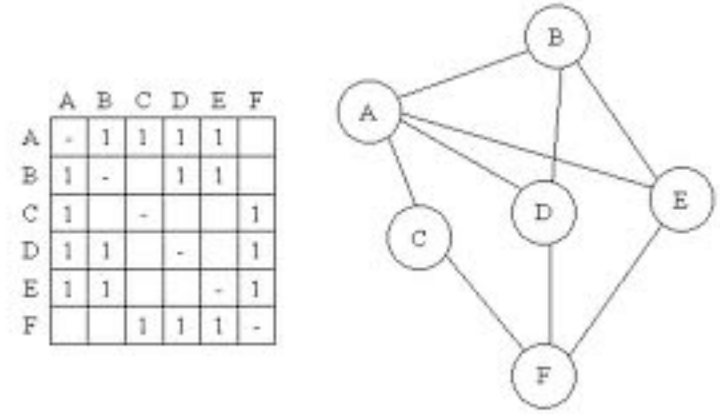
2. **Edit events**
(time stamps for all changes in the wikipedia pages)

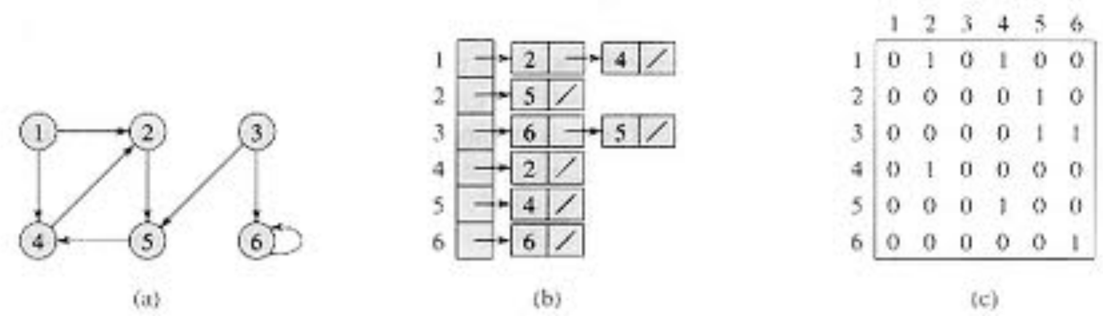


- **filtering, resampling**
- **feature extraction** (peak detection)
- creation of (non)-overlapping **episodes** or (sliding) windows
- creation of time series pairs for **cross-correlation or event synchronisation**

preprocessing, calculation on single records

Map-Reduce / UDF





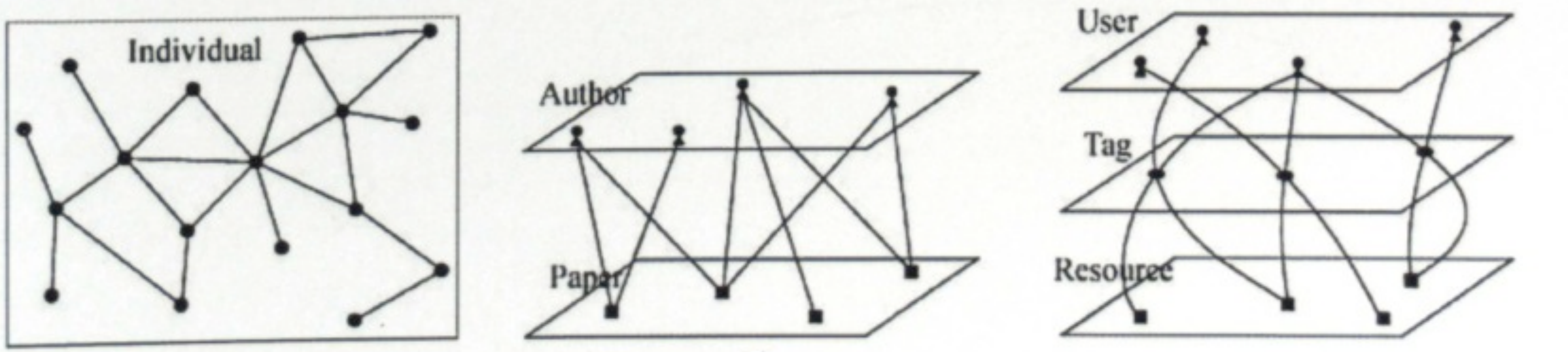
- Graph Analysis
  - If network data is prepared as an **adjacency list** or an **adjacency matrix**, tools like Giraph or Mahout work well.

But: only if the appropriate data strcutures and Input-Format Readers exist.

TYPES OF NETWORKS

a) **unipartite network**, one type of nodes and links
b) **bipartite network**, one type of connections
c) **hypergraph**, one link relates more than two nodes

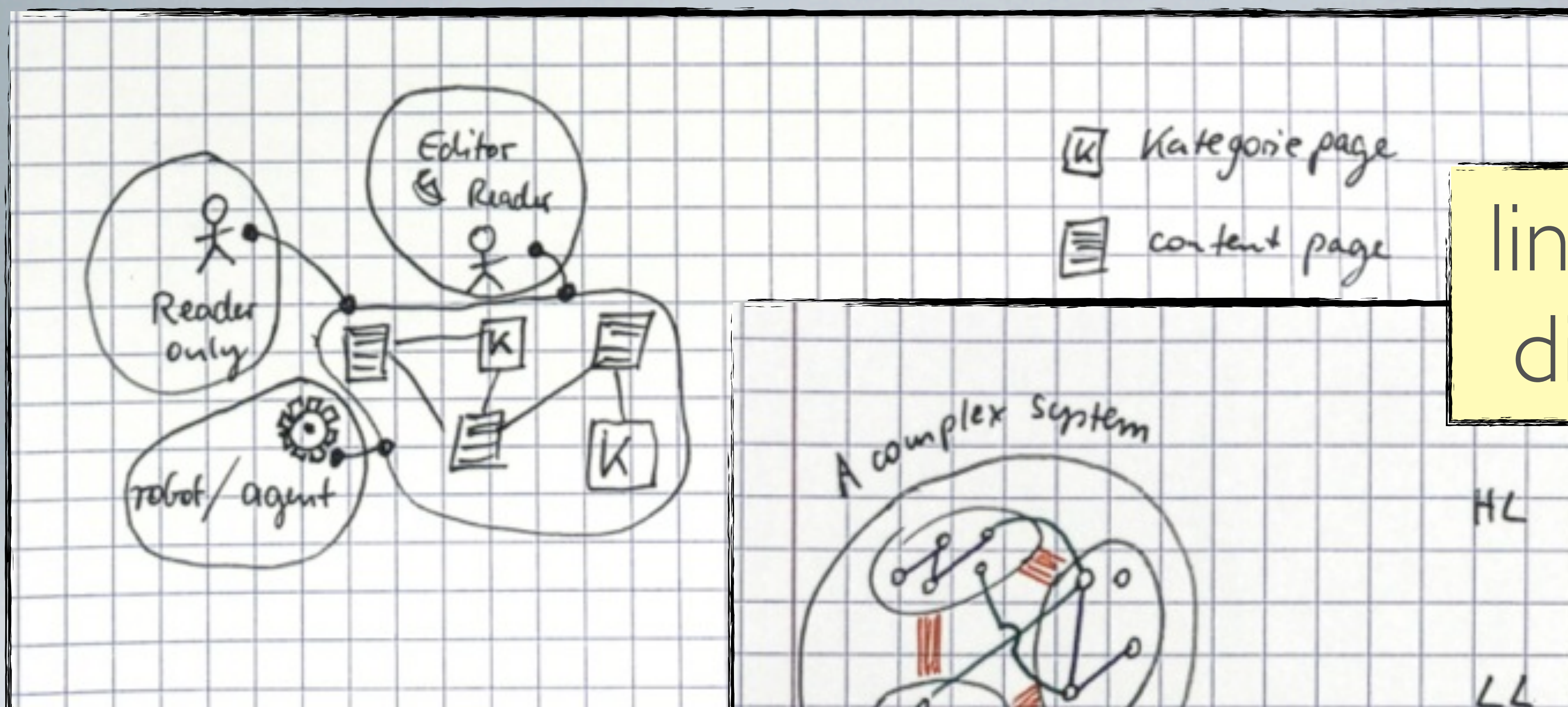
links of just one single type



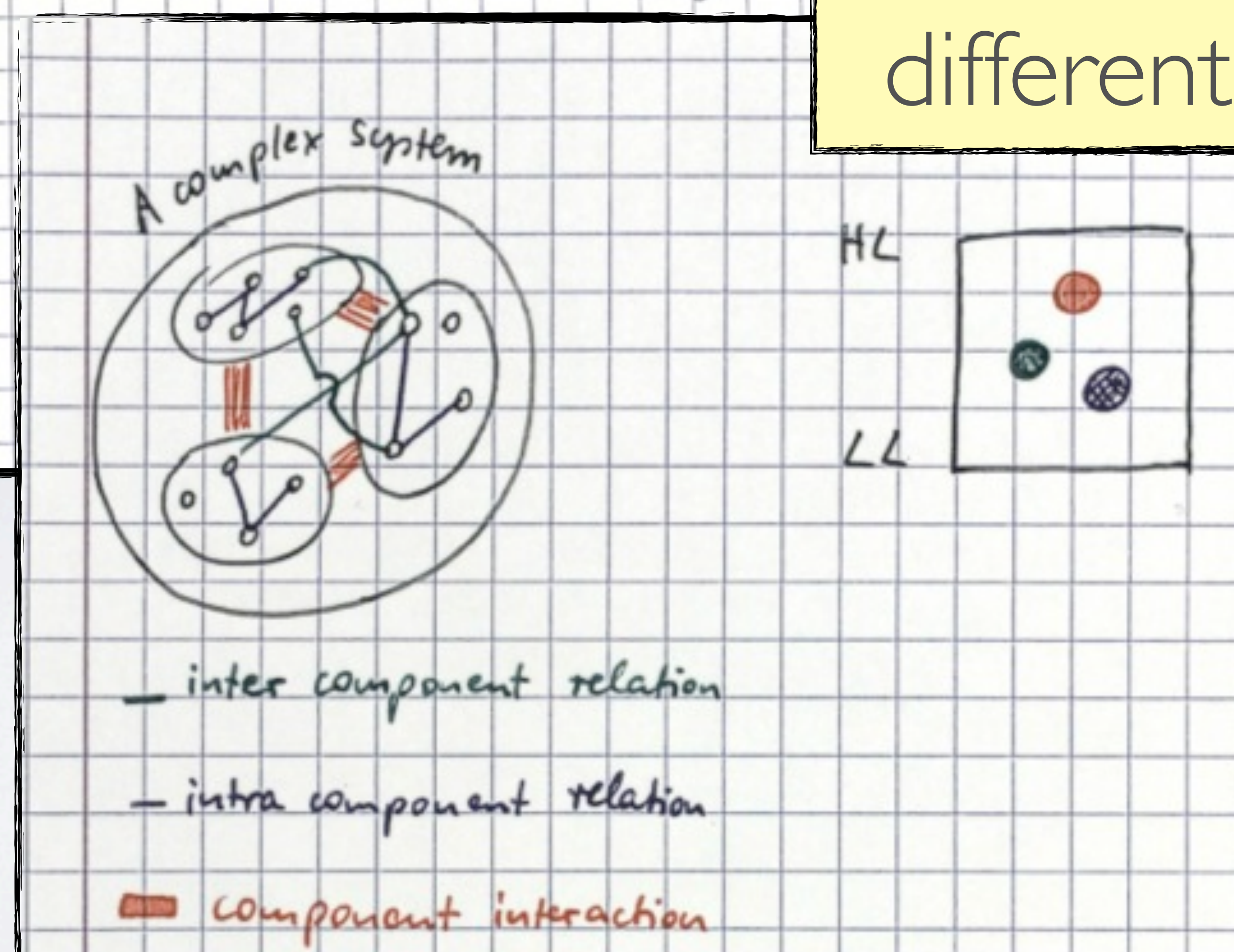
(a) (b) (c)

TYPES OF NETWORKS

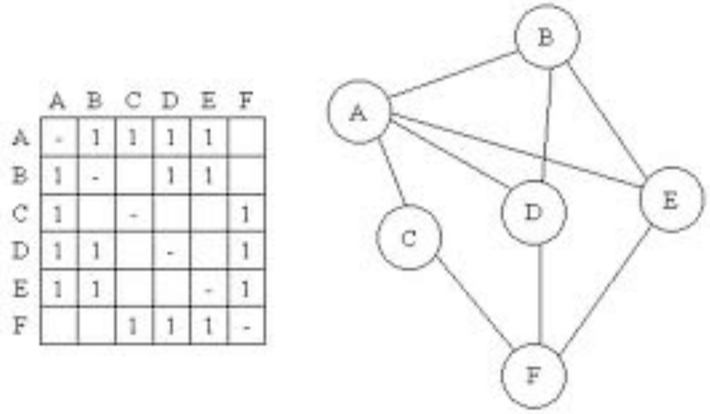
MULTIPLEX NETWORKS

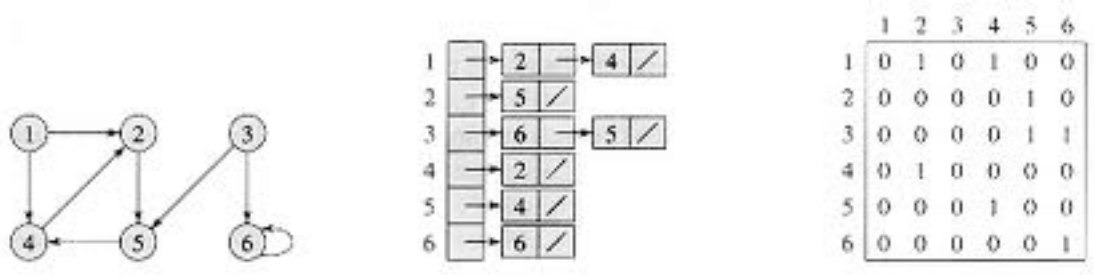


links of multiple different types



HIERARCHICAL NETWORKS





(a) (b) (c)

- creation of adjacency matrix is not trivial
- adjacency matrix is an ineffiecient format
- files stored in HDFS are read only and can not be changed

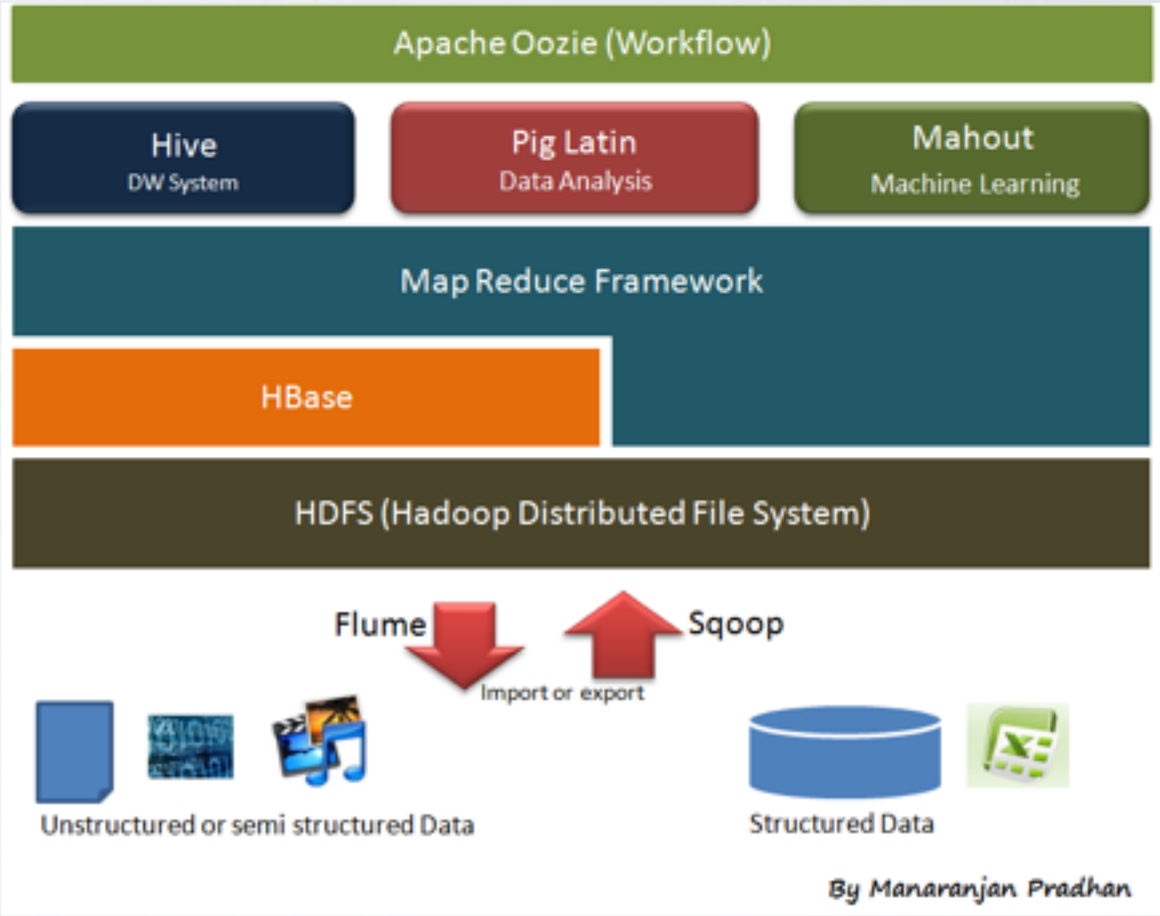
- store dynamic edge / node properties in HBase
- aggregate relevant data to network snapshots
- store intermediate results back to HBase and preprocess this data in a following utility-step

- **Graph Analysis**
  - Large scale raw data sets have to be stored and processed in a scalable distributed system.

  - **How to organize node/edge properties?**
  - **How to deal with time dependent properties?**
  - **How to calculate link properties on the fly?**

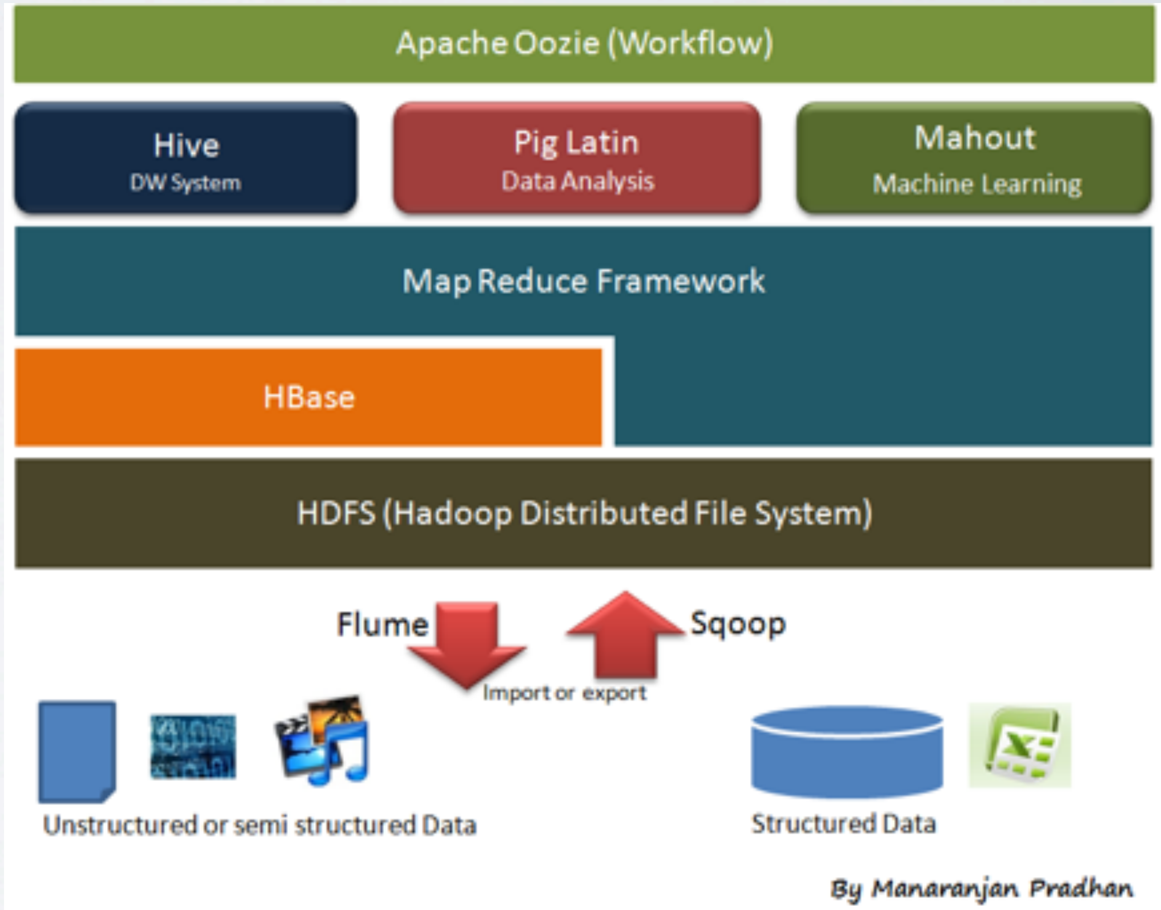
RECAP: WHAT IS HADOOP ?

- Distributed platform to store and process massive amounts of data in parallel
- Implements Map-Reduce paradigm on top of **H**adoop **D**istributed **F**ile **S**ystem.



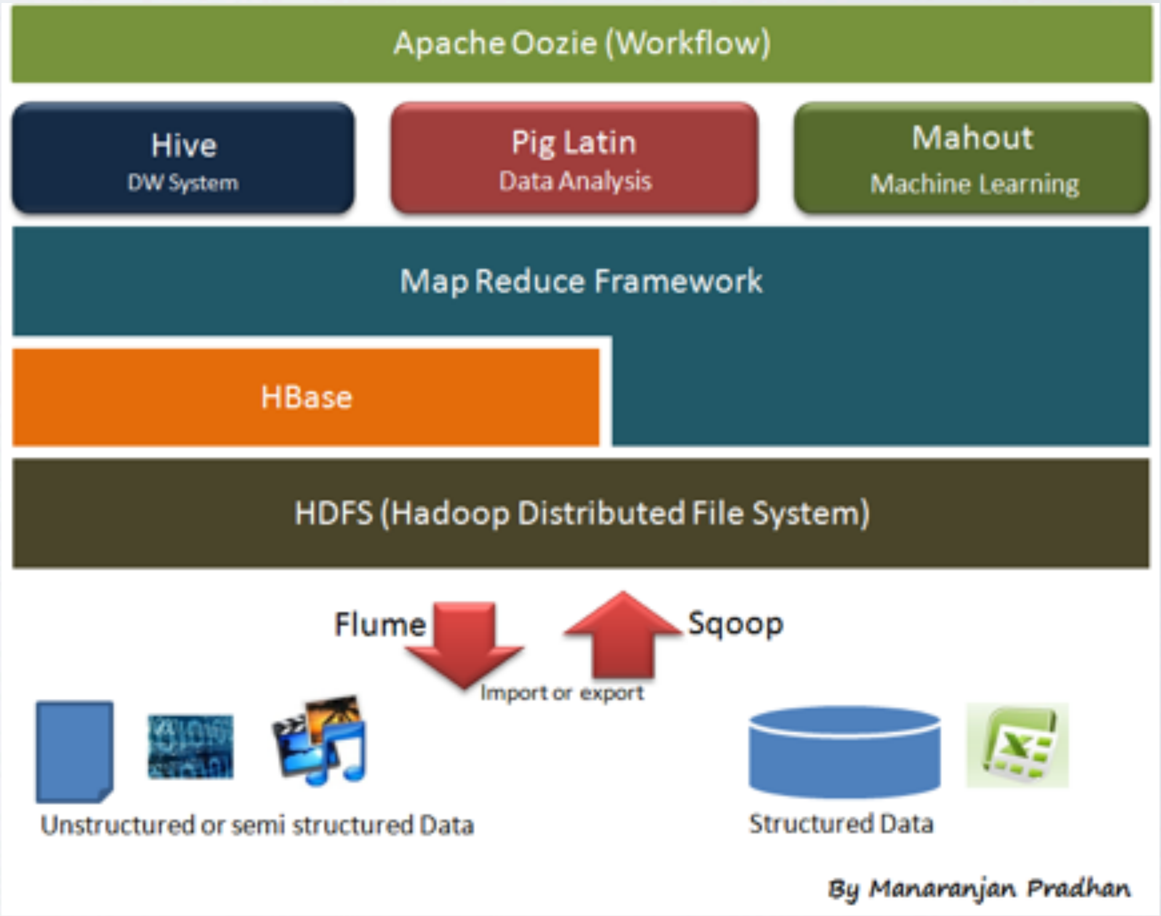
RECAP: **WHAT IS HADOOP ?**

- Distributed platform to store and process massive amounts of data in parallel
- Implements Map-Reduce paradigm on top of Hadoop Distributed File System.
- Map-Reduce : JAVA API to implement a map and a reduce phase
  - **Map phase** uses *key/value* pairs,
  - **Reduce phase** uses *key/value-list* pairs



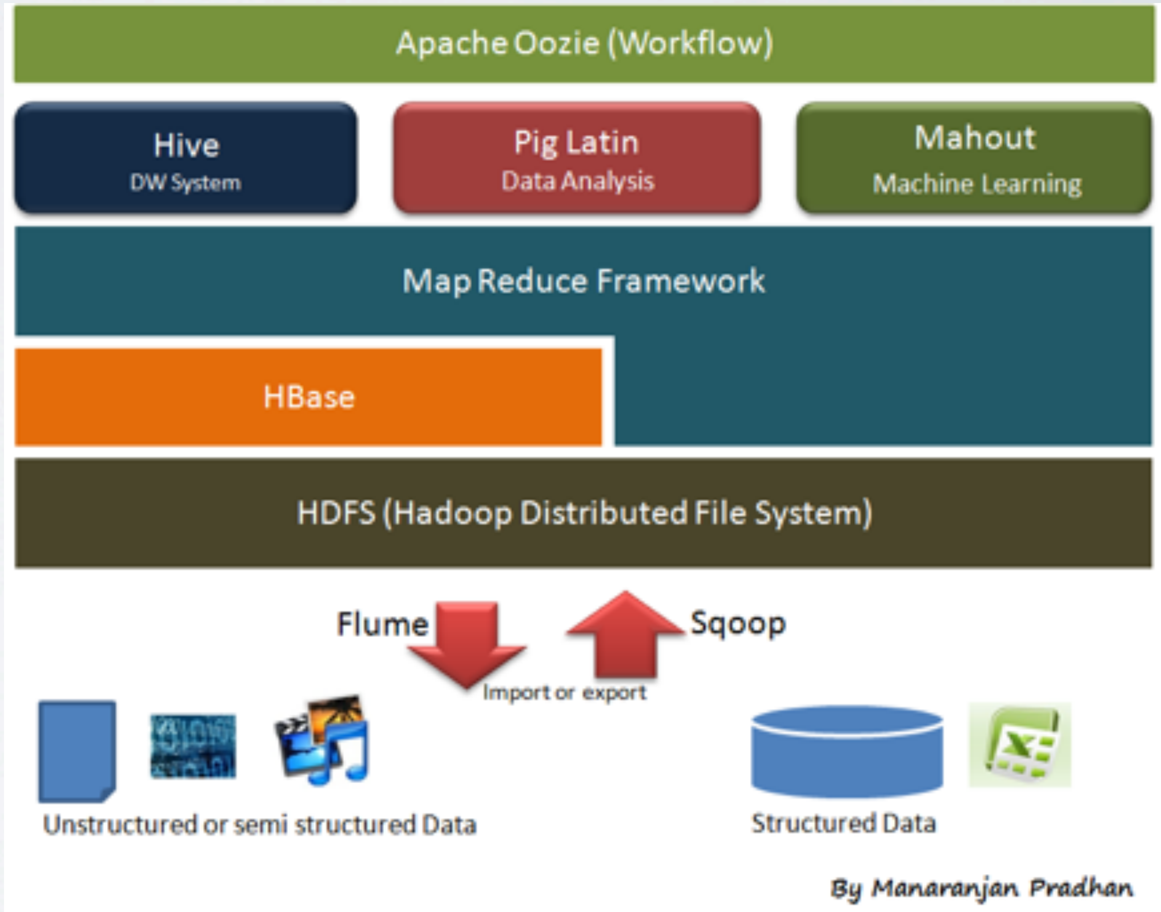
RECAP: WHAT IS HADOOP ?

- Distributed platform to store and process massive amounts of data in parallel
- Implements Map-Reduce paradigm on top of Hadoop Distributed File System.
- Map-Reduce : JAVA API to implement a map and a reduce phase
  - **Map phase** uses *key/value* pairs,
  - **Reduce phase** uses *key/value-list* pairs
- HDFS files consist of one or more blocks (distributed chunks of data).



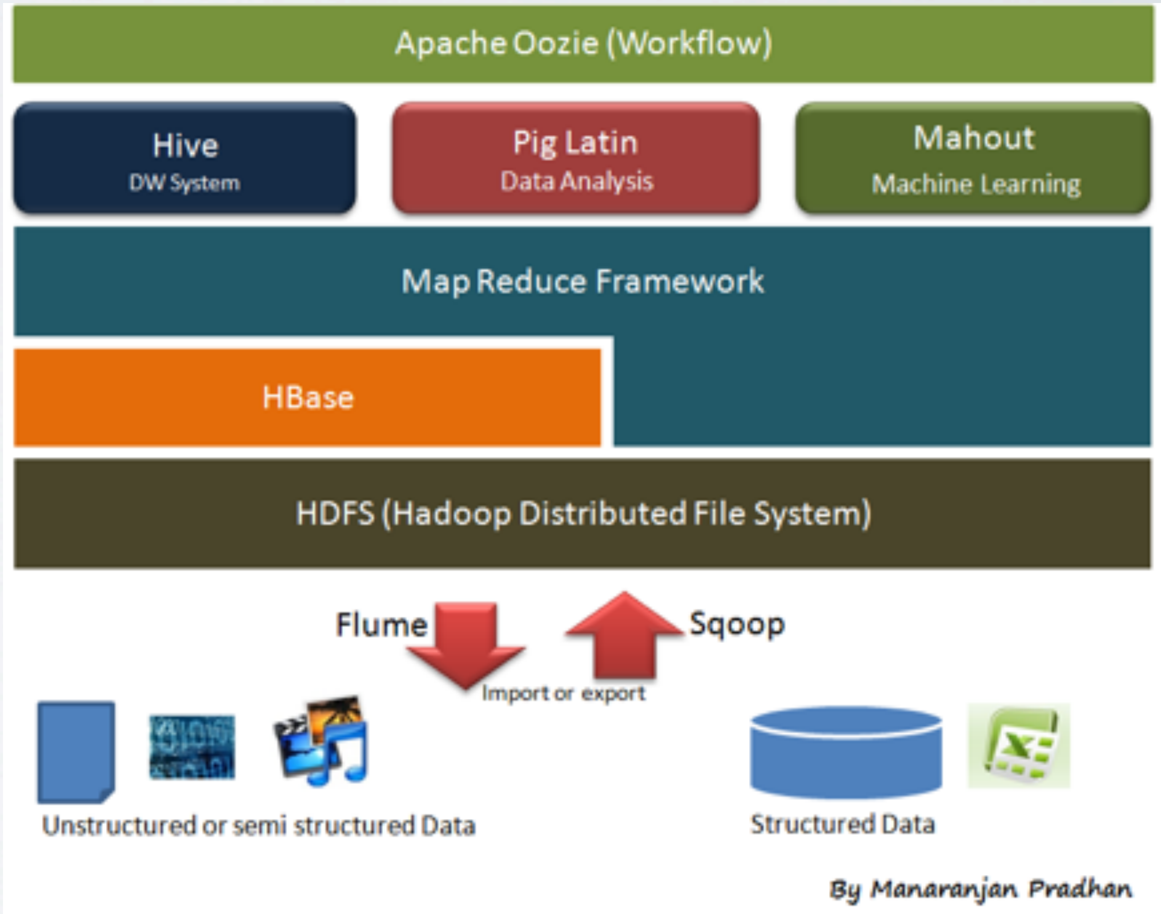
RECAP: **WHAT IS HADOOP ?**

- Distributed platform to store and process massive amounts of data in parallel
- Implements Map-Reduce paradigm on top of Hadoop Distributed File System.
- Map-Reduce : JAVA API to implement a map and a reduce phase
  - **Map phase** uses *key/value* pairs,
  - **Reduce phase** uses *key/value-list* pairs
- HDFS files consist of one or more blocks (distributed chunks of data).
- Chunks are distributed transparently (in background) and processed in parallel.



RECAP: WHAT IS HADOOP ?

- Distributed platform to store and process massive amounts of data in parallel
- Implements Map-Reduce paradigm on top of Hadoop Distributed File System.
- Map-Reduce : JAVA API to implement a map and a reduce phase
  - **Map phase** uses *key/value* pairs,
  - **Reduce phase** uses *key/value-list* pairs
- HDFS files consist of one or more blocks (distributed chunks of data).
- Chunks are distributed transparently (in background) and processed in parallel.
- Using data locality when possible by assigning the map task to a node that contains the chunk locally.



MAP REDUCE: TYPICAL APPLICATIONS

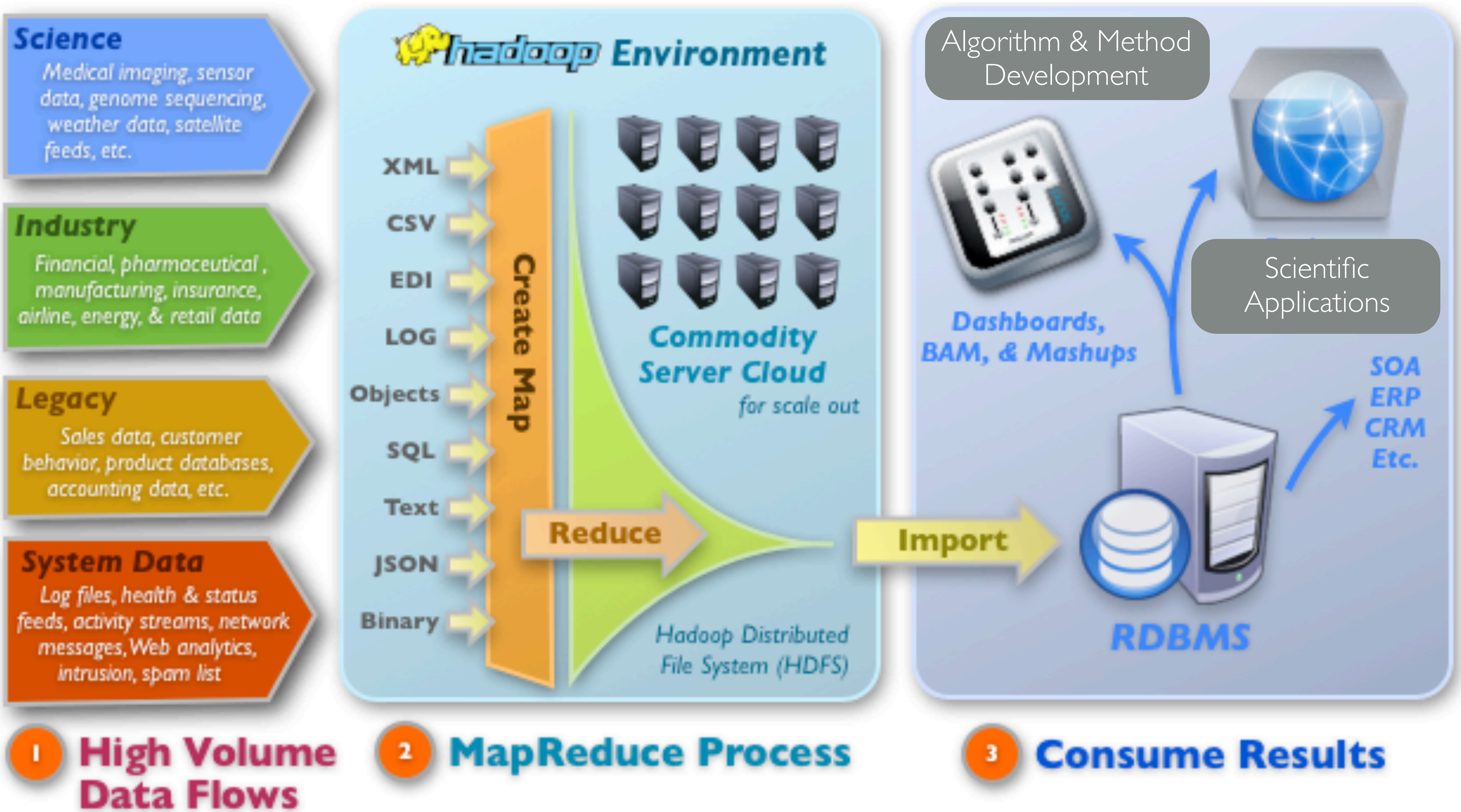
- **Filter**, **group**, and **join operations** on large data sets ...
  - the data set (**or a part of it**)* is streamed and processed in parallel, but usually not in real time

* if partitioning is used

- Algorithms like **k-Means Clustering** (Apache Mahout) or Map-Reduce based implementations of **SSSP** work in **multiple iterations**
  - data is loaded from disk to CPU in each iteration!

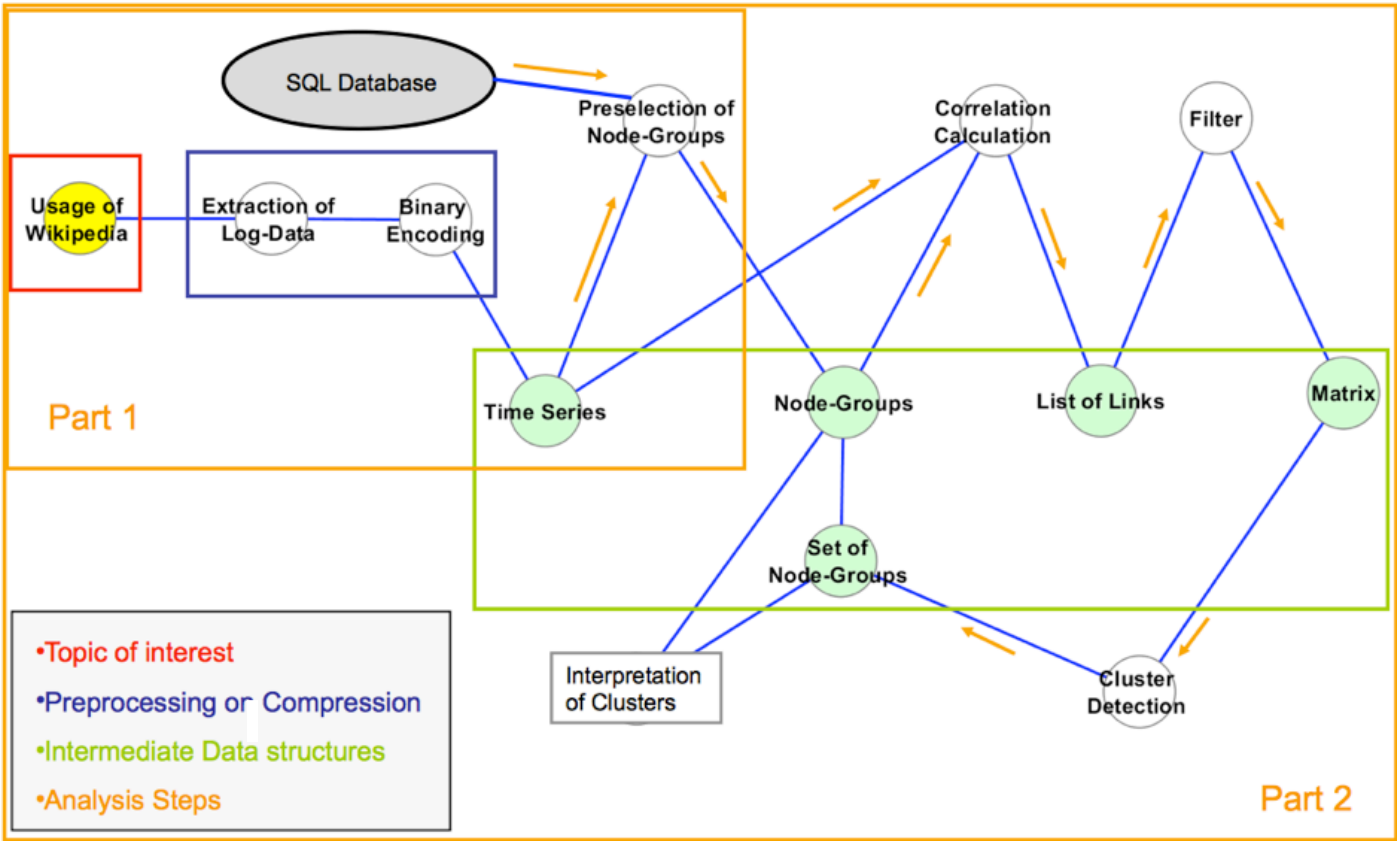
* heavy I/O workload

HADOOP:
Platform for large scale data integration

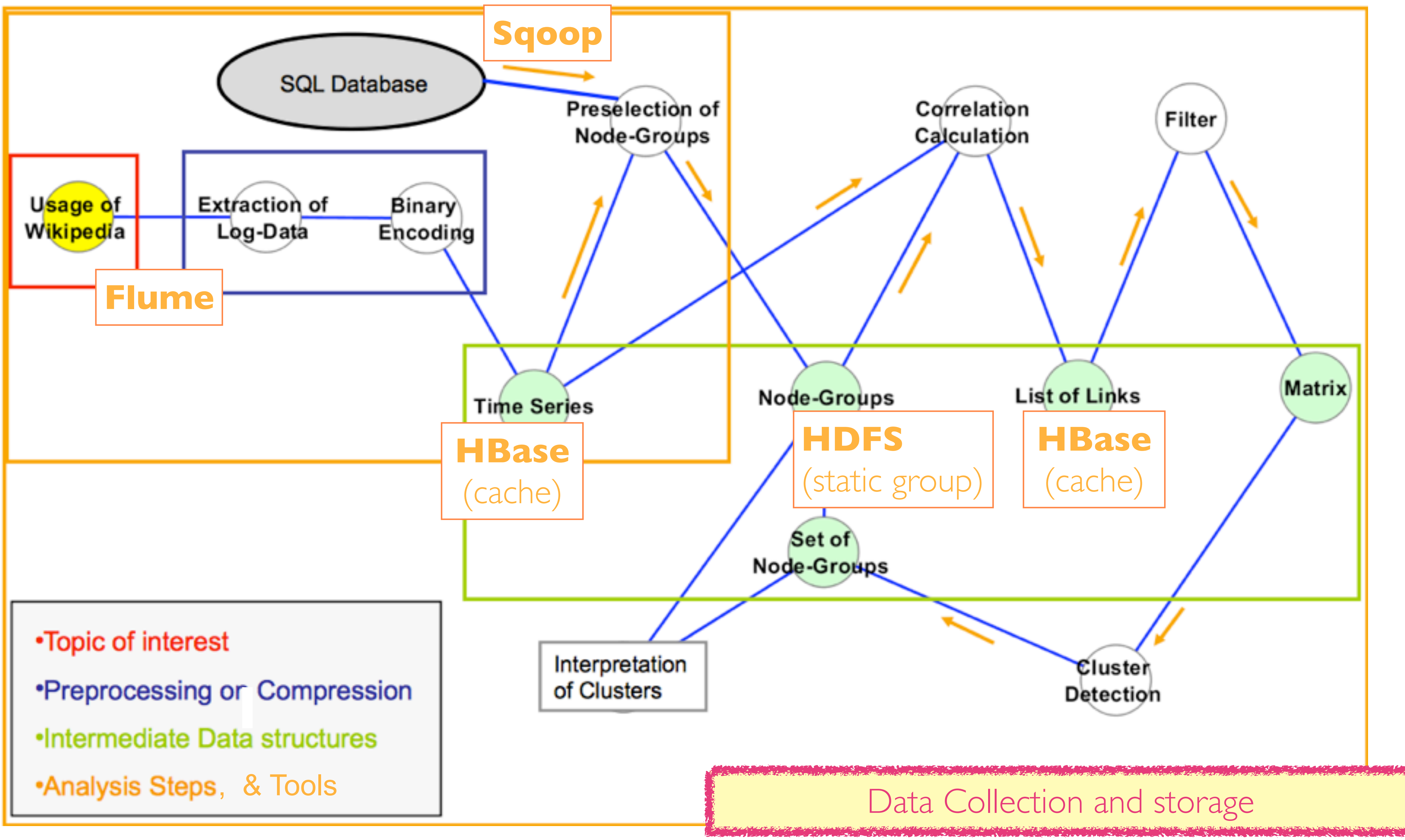


From http://www.ebizq.net/blogs/enterprise

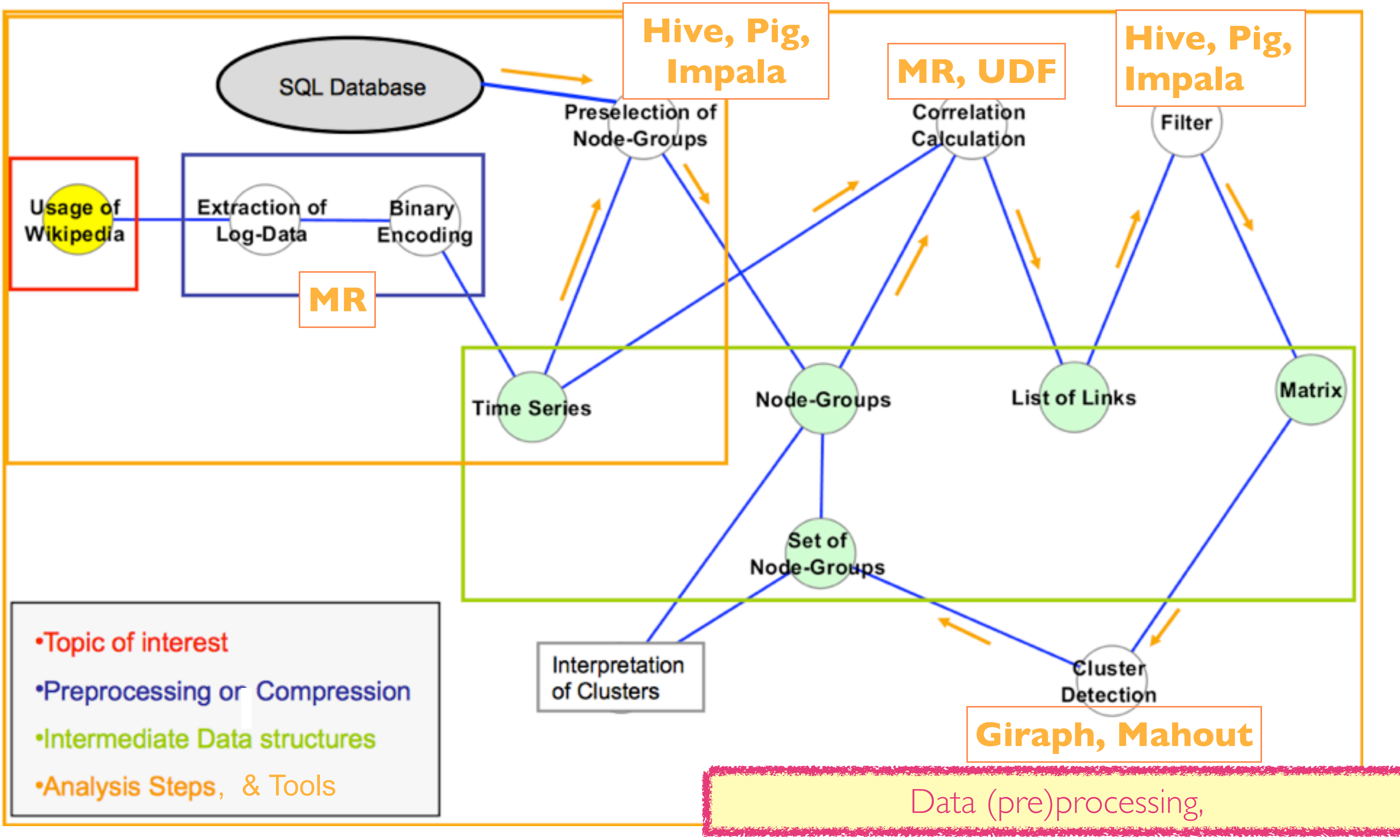
OVERVIEW - DATA FLOW



OVERVIEW - DATA FLOW

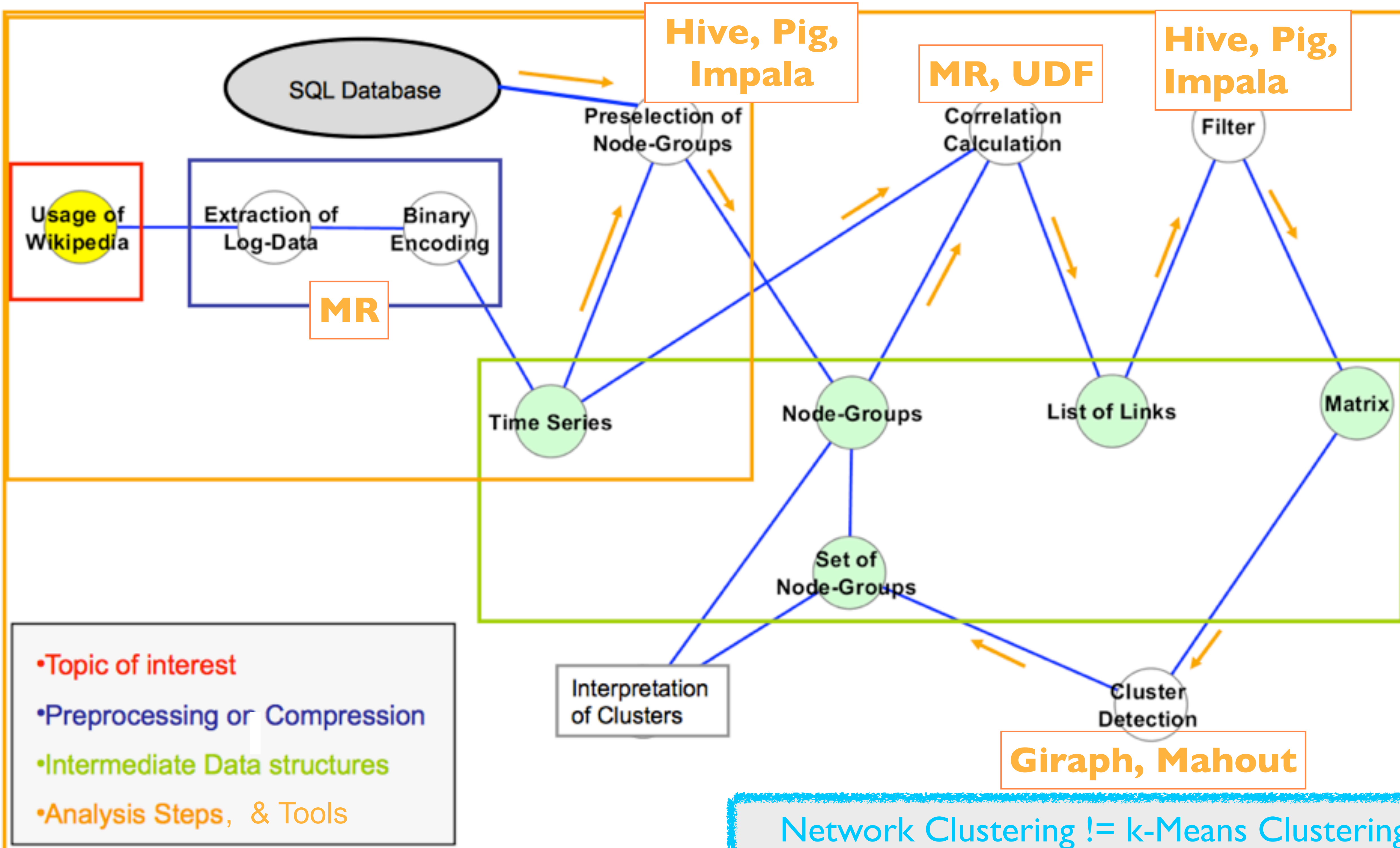


OVERVIEW - DATA FLOW



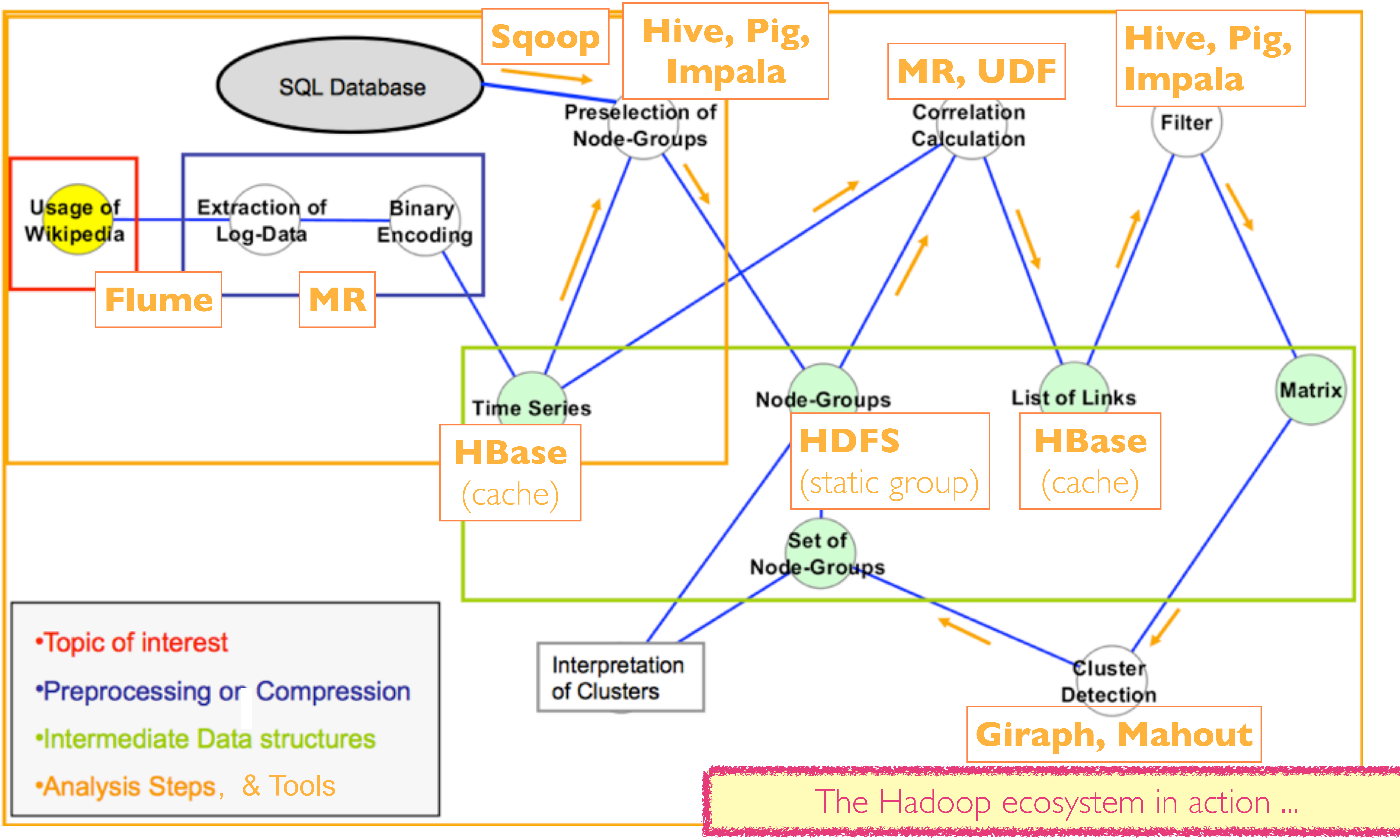
Dienstag, 27. August 13

OVERVIEW - DATA FLOW

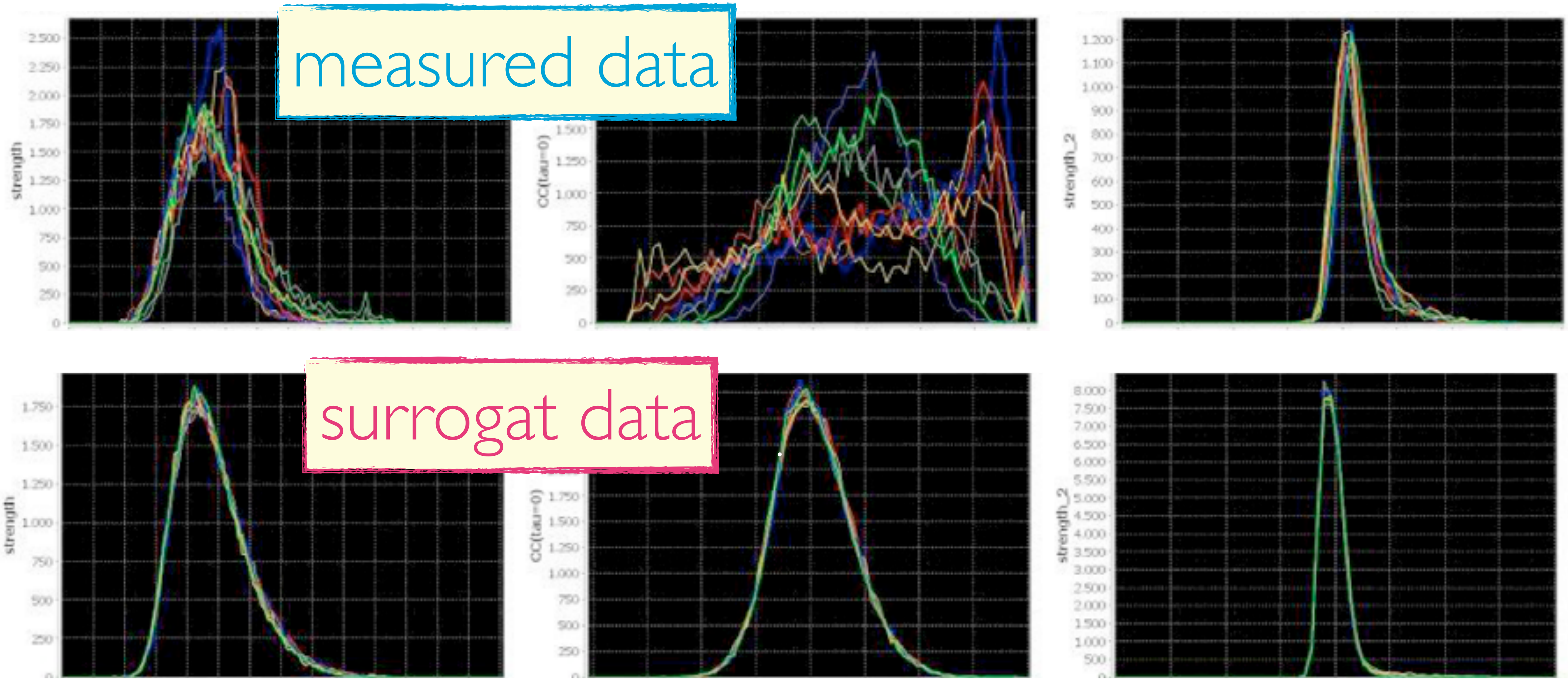


Network Clustering != k-Means Clustering

OVERVIEW - DATA FLOW



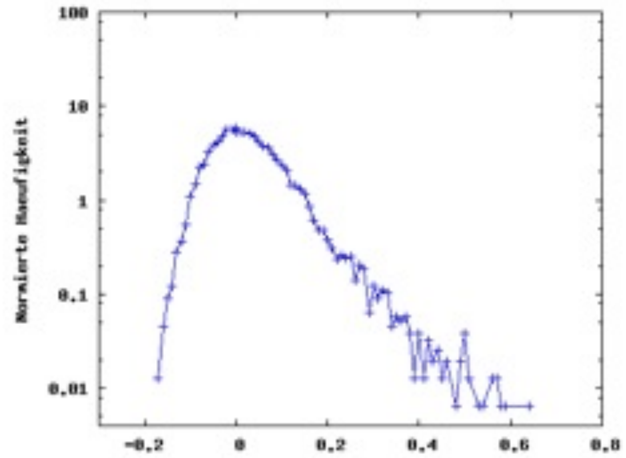
RESULTS: CROSS-CORRELATION

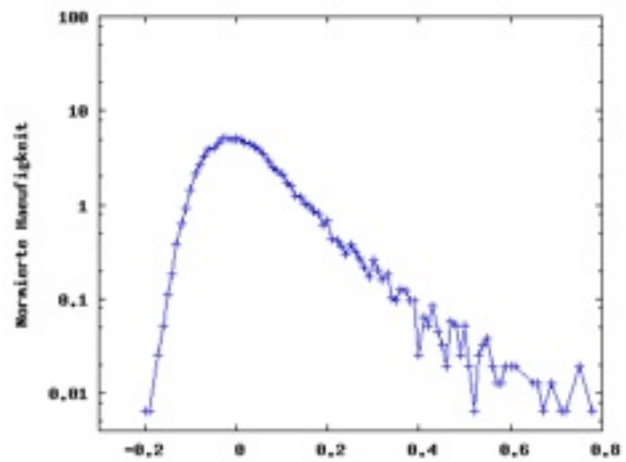


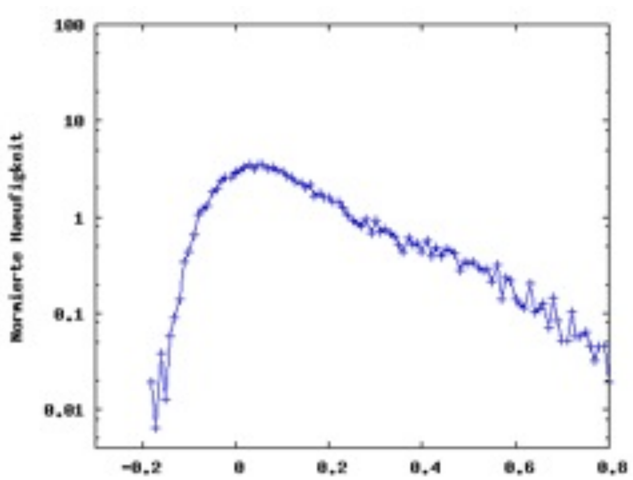
Distribution of cross-correlation coefficients for pairs of access-rate time series of Wikipedia pages (top) compared to surrogat data (bottom) - 100 shuffled configurations are considered

EVOLUTION OF CORRELATION NETWORKS

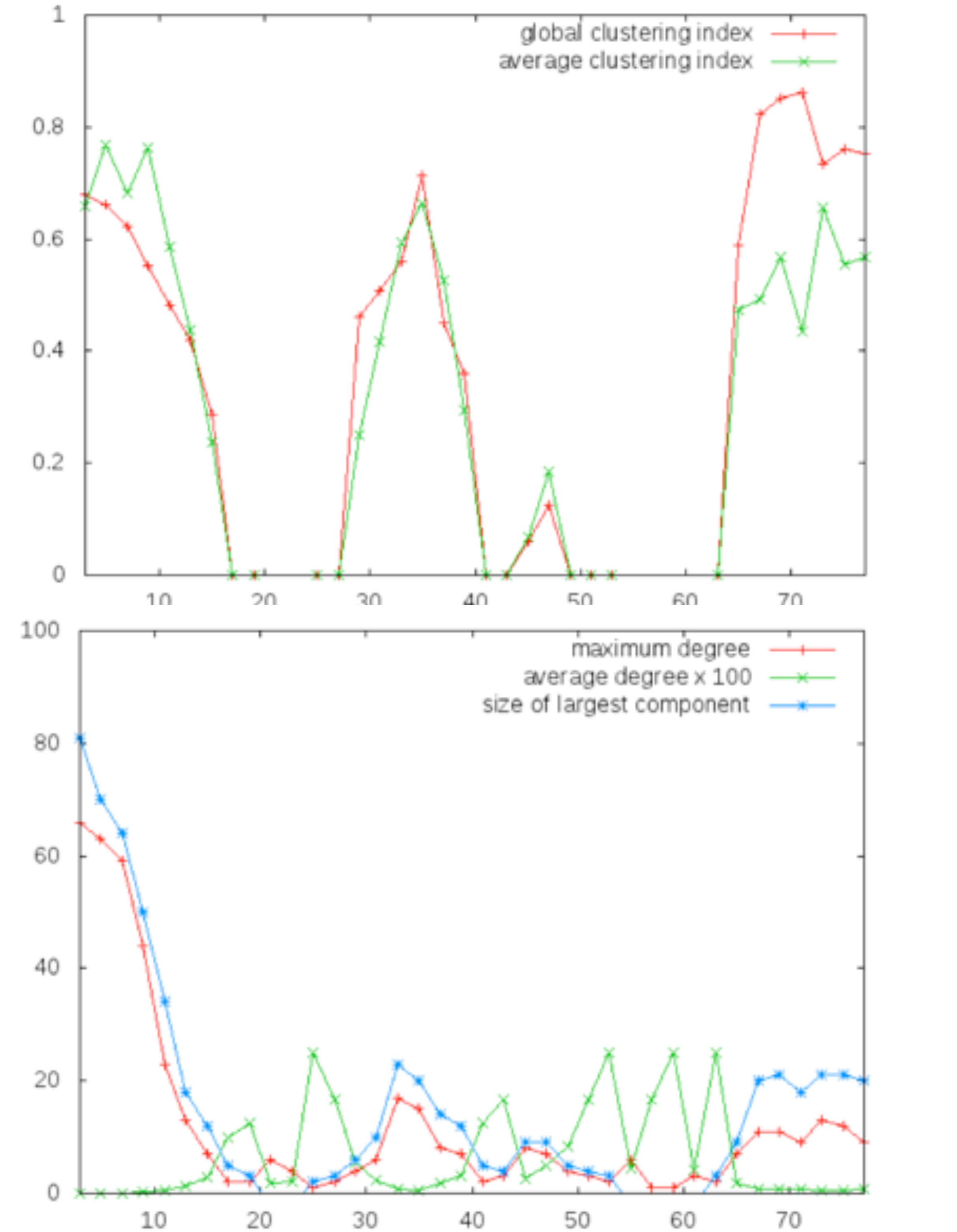
time





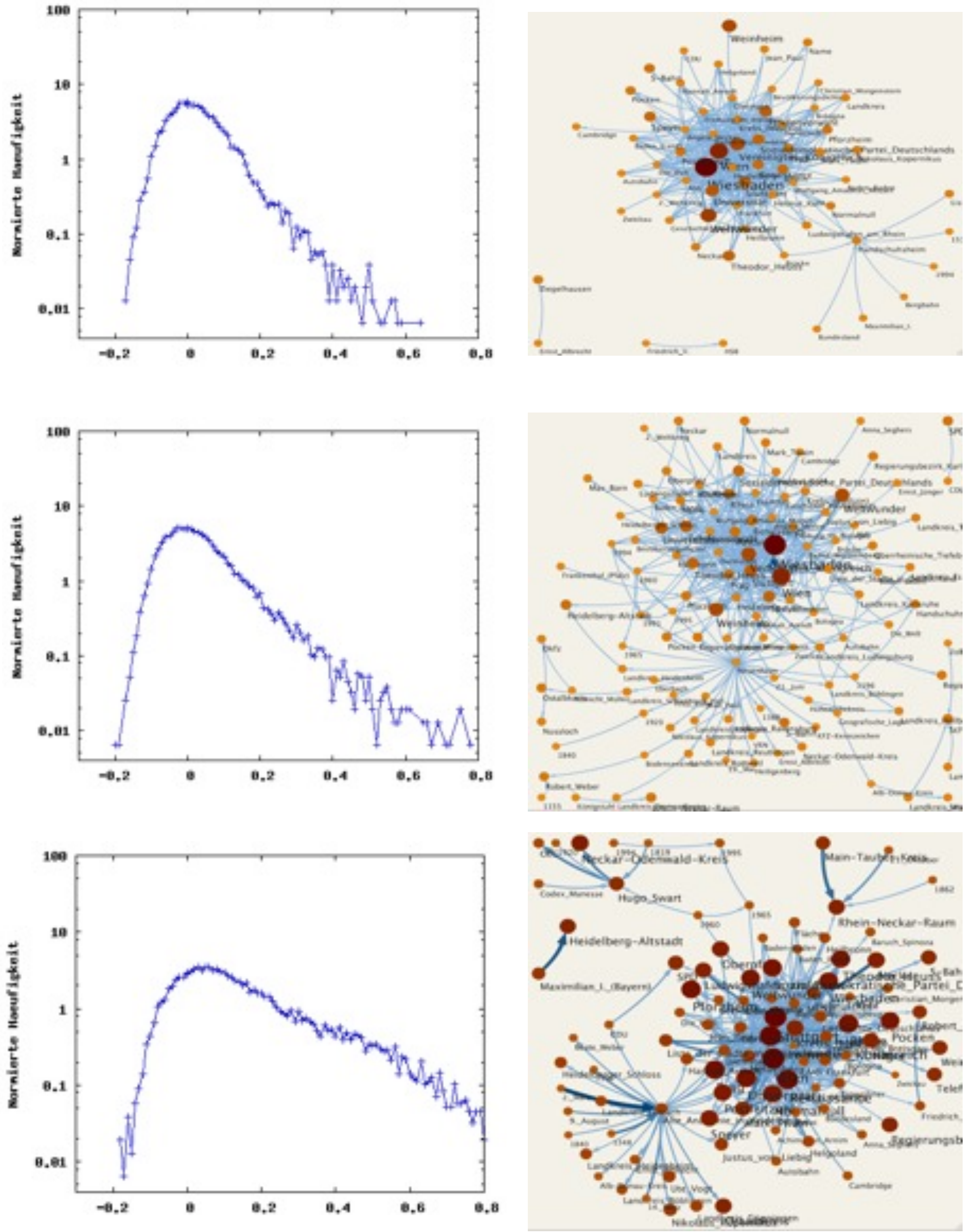


time

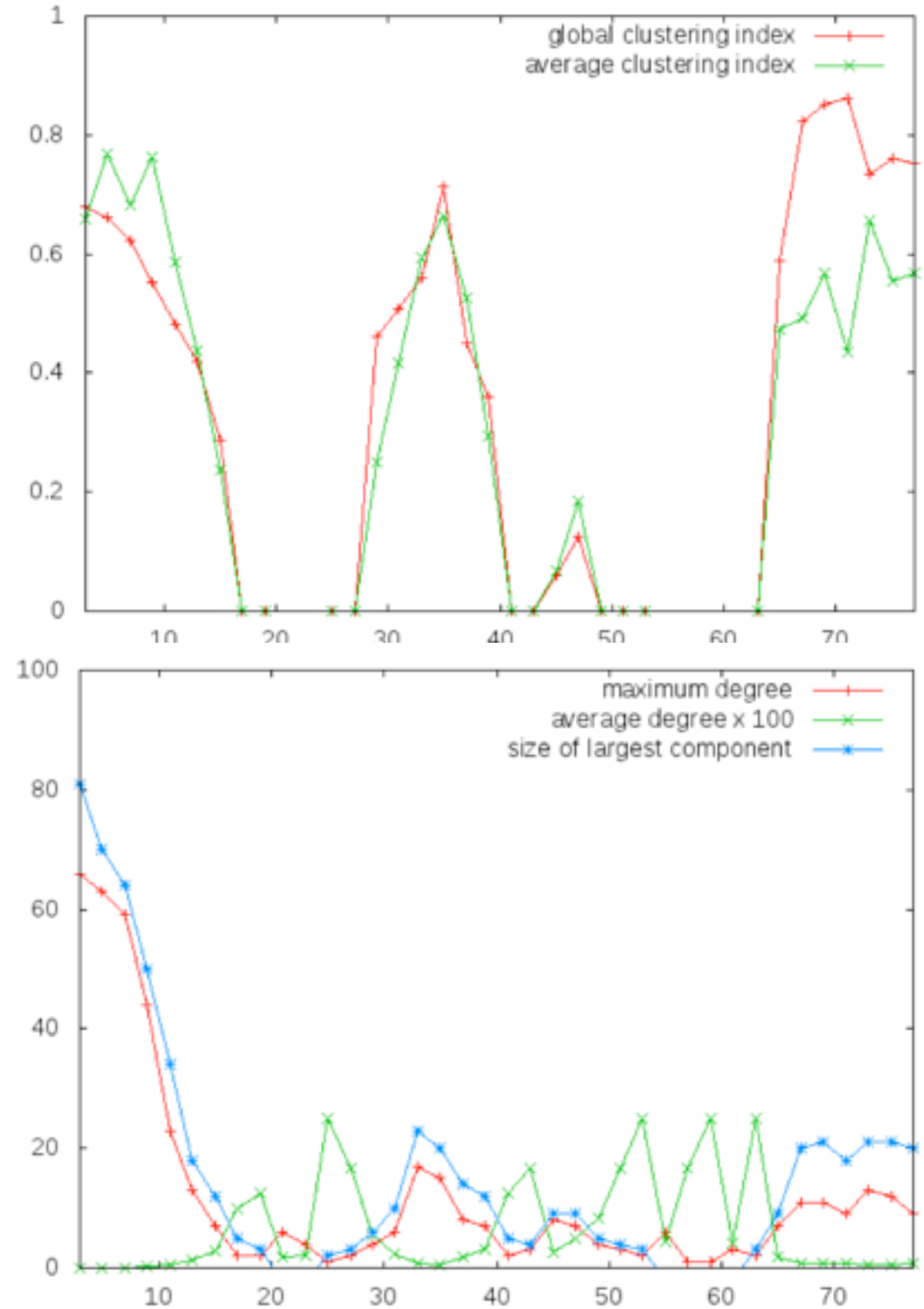


EVOLUTION OF CORRELATION NETWORKS

time

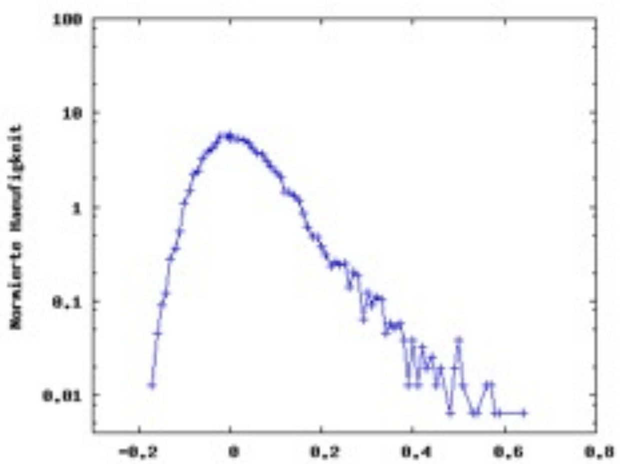


time

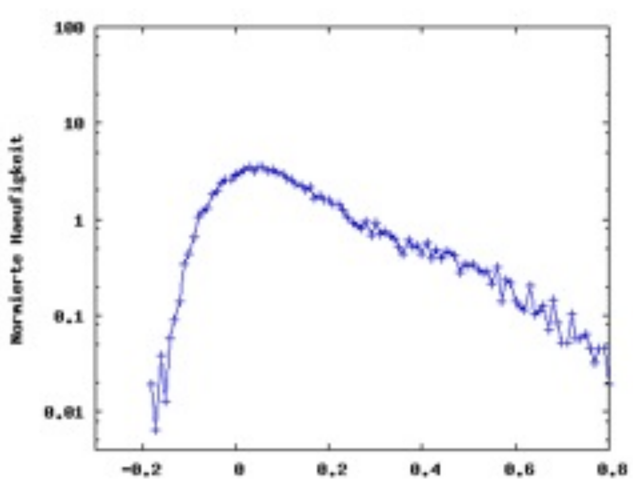


EVOLUTION OF CORRELATION NETWORKS

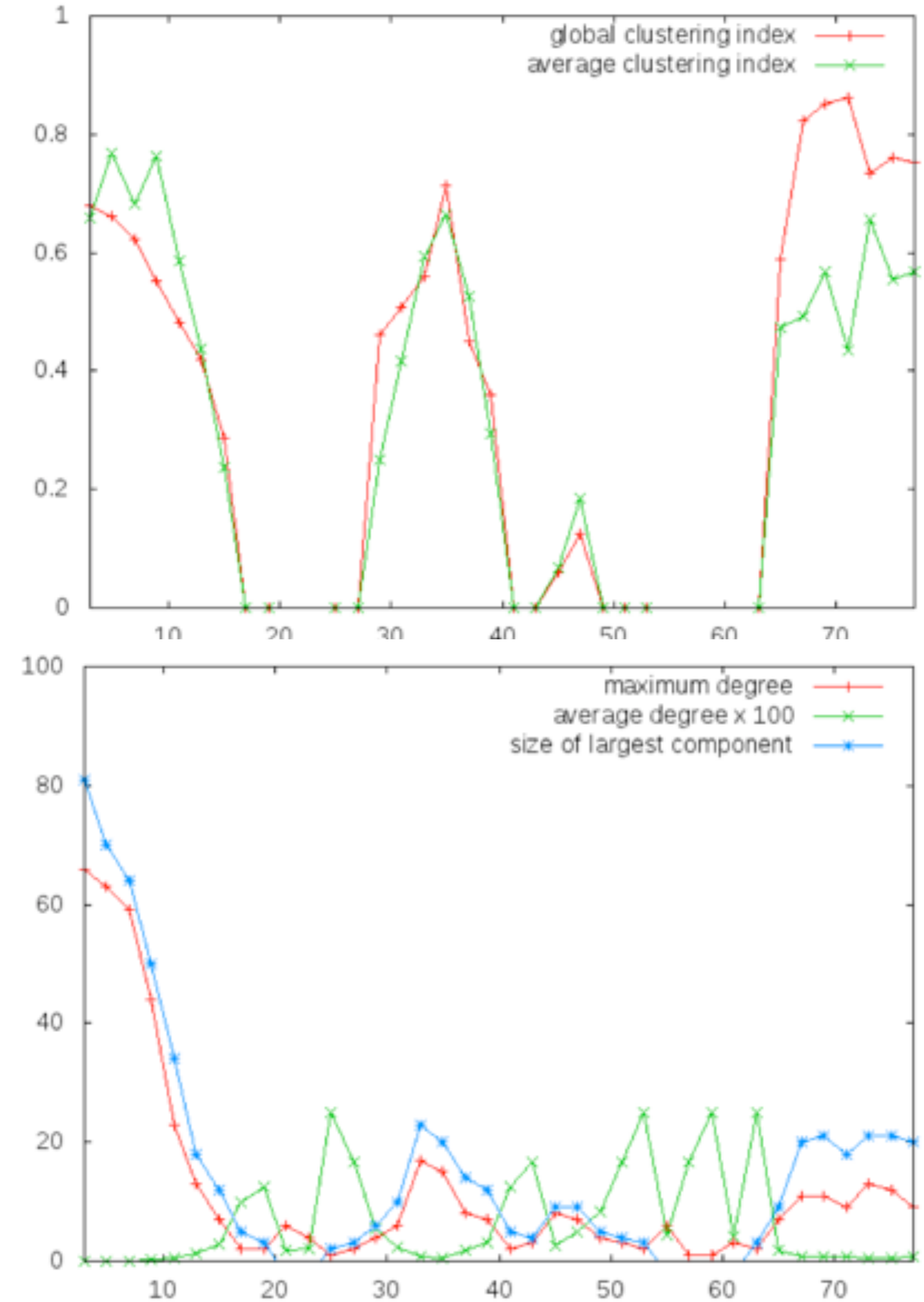
time



Obviously, the **distribution** of link strength values **changes in time**, but only a calcultion of **structural properties** of the underlying network allows a detailed view on the **dynamics if the system**.

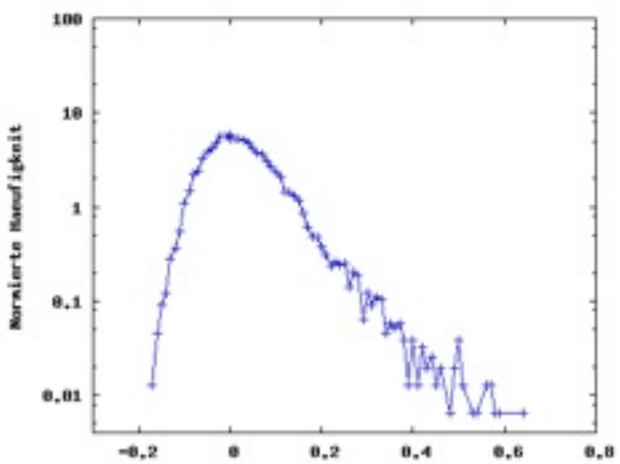


time

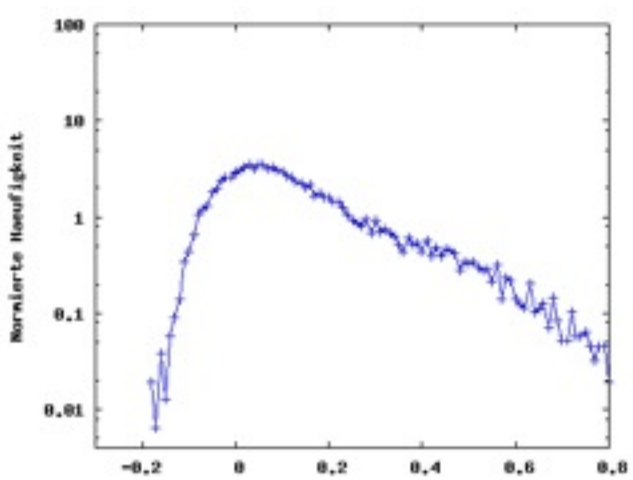


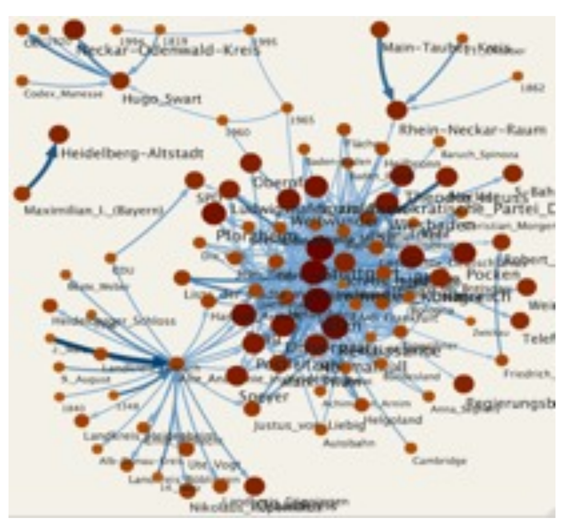
EVOLUTION OF CORRELATION NETWORKS

time

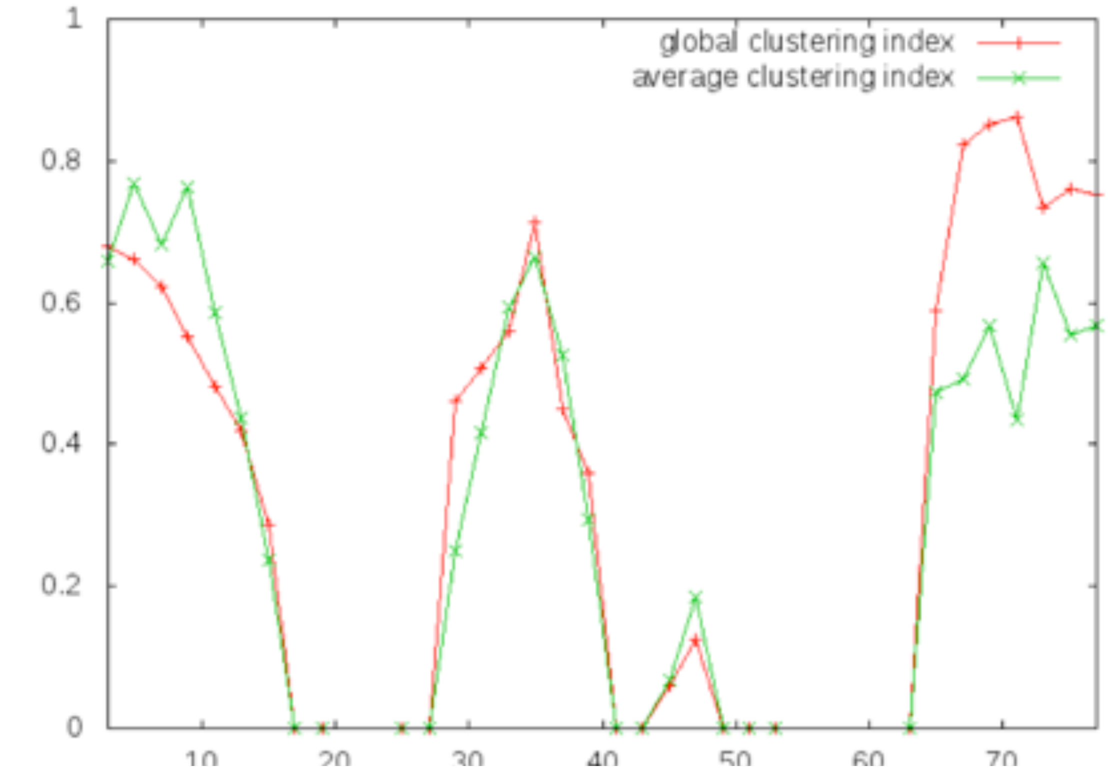


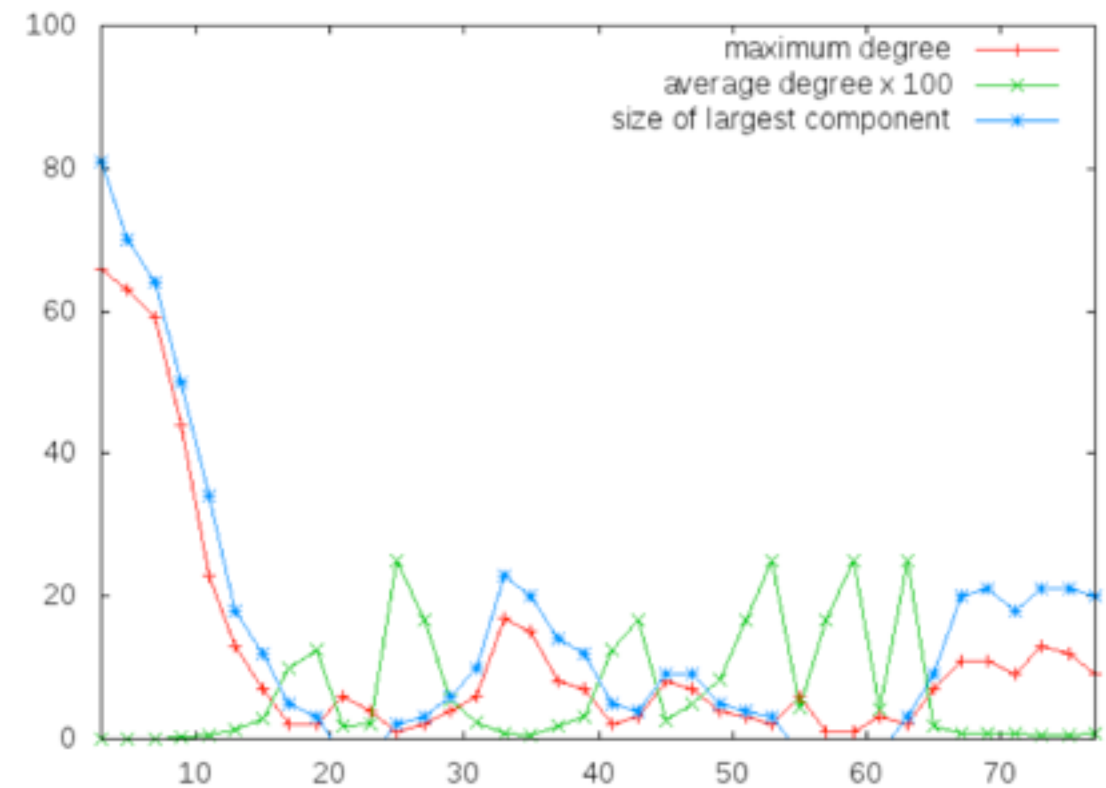
Obviously, the **distribution** of link strength values **changes in time**, but only a calcultion of **structural properties** of the underlying network allows a detailed view on the **dynamics if the system**.





time





EXAMPLES OF RECONSTRUCTED NETWORKS

Spatially embedded correlation networks,
for wikipedia pages of all German cities.

Wikipedia Access Network

Wikipedia Edit Network

RECOMMENDATIONS (I.)

- Create algorithms based on **reusable components!**
- Use or create **stable and standardized I/O-Formats!**
- Do preprocessing, e.g. a **re-organization of unstructured data**, if you have to process the data many times.
- Collect **event data in HBase** and create **Time-Series Buckets** for advanced procedures, maybe on a subset of the data.
- Store intermediate data (e.g. time dependent properties) in **HBase**, close to the raw data, and **allow random access**.

RECOMMENDATIONS (2.)

- Consider Design Patterns
  - Partitioning vs. Binning
  - Map-Side vs. Reduce-Side Joins
- Use **B**ulk **S**ynchronuos **P**rocessing for graph processing instead of Map-Reduce, or even a combination of both.
- In classical programming: (**and also in Hadoop !!!**) find **good data representation** to find **good algorithms.**
- **Think about access patterns:** streaming vs. random access

REFERENCES

[1] M. Kämpf, S. Tismer, J. W. Kantelhardt, and L. Muchnik; **Burst event and return interval statistics in Wikipedia access and edit data**, submitted to Physica A (2011).

[2] L. Mitchell, M.E. Cates; **Hawkes Process as a model of social interactions : a view on video dynamics**; J. Phys. A: Math. Theor. **43** (2010) 045101.

[3] R. Crane, D. Sornette; **Robust dynamic classes revealed by measuring the response function of a social system**; PNAS **105** (2008) 15649-15653.

[4] C. K. Peng, S. V. Buldyrev, S. Havlin, M. Simons, H. E. Stanley, and A. L. Goldberger; **Mosaic organization of DNA nucleotides**; Phys. Rev. E **49** (1994) 1685.

[5] J.W. Kantelhardt, S.A. Zschiegner, E. Koscielny-Bunde, S. Havlin, A. Bunde, and H.E. Stanley**; Multifractal detrended fluctuation analysis of nonstationary time series**; Physica A **316** (2002) 87.

[6] J.F. Eichner, J.W. Kantelhardt, A. Bunde, and S. Havlin; **Statistics of return intervals in long-term correlated records**; Phys. Rev. E **75** (2007) 011128.

[7] A. Bunde, J.F. Eichner, J.W. Kantelhardt, and S. Havlin; **Long-Term Memory: A Natural** Mechanism for the Custering of Extreme Events; Phys. Rev. Lett. **94** (2005) 048701.

[8] M. Rosevall, D. Axelsson, C.T. Bergstrom; **The map equation**, Eur. Phys J. Special Topics **178** (2009) 13-23.

[9] J.F. Donges, Y. Zuo, N. Marvan and J. Kurths; **Complex networks in climate dynamics : Comparing linear and non-linear network construction methods**, Eur. Phys J. Special Topics **174** (2009) 157-179.

[10] R.Q. Quiroga, T. Kreuz, and P. Grassberger;**Event synchronization: A simple and fast method to measure synchronicity and time delay patterns**, Phys. Rev. E **66** (2002) 041904.

[11] N. Malik, B. Bookhagen, N. Marwan, and J. Kurths; **Analysis of spatial and temporal extreme monsoonal rainfall over South Asia using complex networks**, Clim Dyn (in press 2011), DOI 10.1007/s00382-011-1156-4

[12] S.V. Buldyrev, R. Parshani, G. Paul, H.E. Stanley, and S. Havlin; **Catastrophic cascade of failures in interdependent networks**; Nature **464** (2010) 1025-1028.

[13] S. Havlin, D.Y. Kenett, E. Ben-Jacob, A. Bunde, R. Cohen, H. Hermann, J. W. Kantelhardt, J. Kertesz, S. Kirkpatrick, J. Kurths, Y. Portugali, and S. Solomon; **Challenges of network science: Applications to infrastructures, climate, social systems and economics**, Eur. Phys. J. ST (in print, 2012).

MANY THANKS !!!

Apache

Wikimedia Foundation

hadoop

MANY THANKS !!!

- to the audience, here in Karlsruhe!
- to my supervisor and collaborators at MLU:
  - PD Dr. Jan W. Kantelhardt, Berit Schreck, Arne Böcker
- to my colleagues at Cloudera, Inc.
  - Kai Voigt, Glynn Durham, and Tom Wheeler


cloudera
Ask Bigger Questions