IDEALEM
Implementation of Dynamic Extensible Adaptive Locally Exchangeable Measures

Scientific Data Management Research Group
Computational Research Division
Lawrence Berkeley National Laboratory

Motivation/Observations

- **Motivation**
  - Large streaming data needs a lot of storage.
  - Statistical analysis is needed on big data.
  - Exact compression of big streaming data is intractable, in general.
    - Alternative: Linear random sampling, e.g. 1 out of 1000 records
      - It is not scalable for high-rate multiple streaming data
      - There is no guarantee of reflecting the underlying data distribution
- **Observations**
  - Large streaming data tend to show redundant data patterns.
  - Many conventional statistical methods are based on a specific assumption (exchangeability).

IDEALEM: New Perspective on Data Compression

- **IDEALEM (Implementation of Dynamic Extensible Adaptive Locally Exchangeable Measures)**
- **Relaxing order of values opens up new horizon on data compression**
  - Information loss due to compression has been generally measured by Euclidean distance (L^2-norm) between original data and reconstructed data with MSE/SNR criteria
    - High entropy (nearly random) data and floating-point values are hard to compress
  - Limitation: order of values not preserved
    - Is the order of values really important?
    - Devices such as sensors often measure random fluctuations
    - Exact reproduction of random fluctuations is not necessary

Exchangeable Random Variables

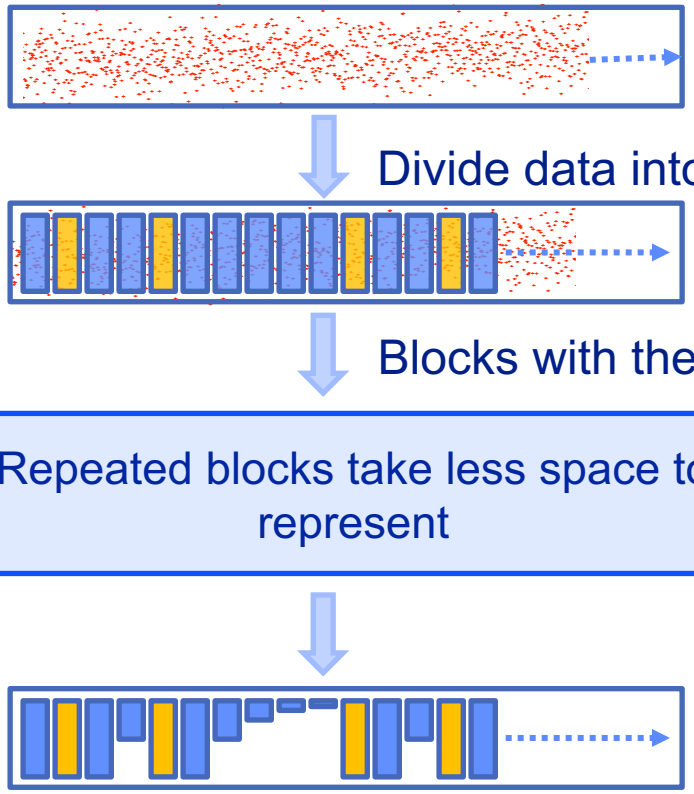
- **Exchangeable RVs: a set of RVs which are interchangeable among others.**

$$P(x_1, \cdots, x_n) = P(x_{\pi(1)}, \cdots, x_{\pi(n)})$$ π: a permutation

- **Exchangeability is already exploited and utilized in many applications such as image & video retrieval and network analysis.**
- **Examples**
  - **Image & video matching: exchangeable image features**
  - **Econometrics: a set of exchangeable portfolio (in risk analysis)**
  - **The Netflix prize: groups of users & groups of movies**

An Illustrative Example of Locally Exchangeable Measures (LEMs)

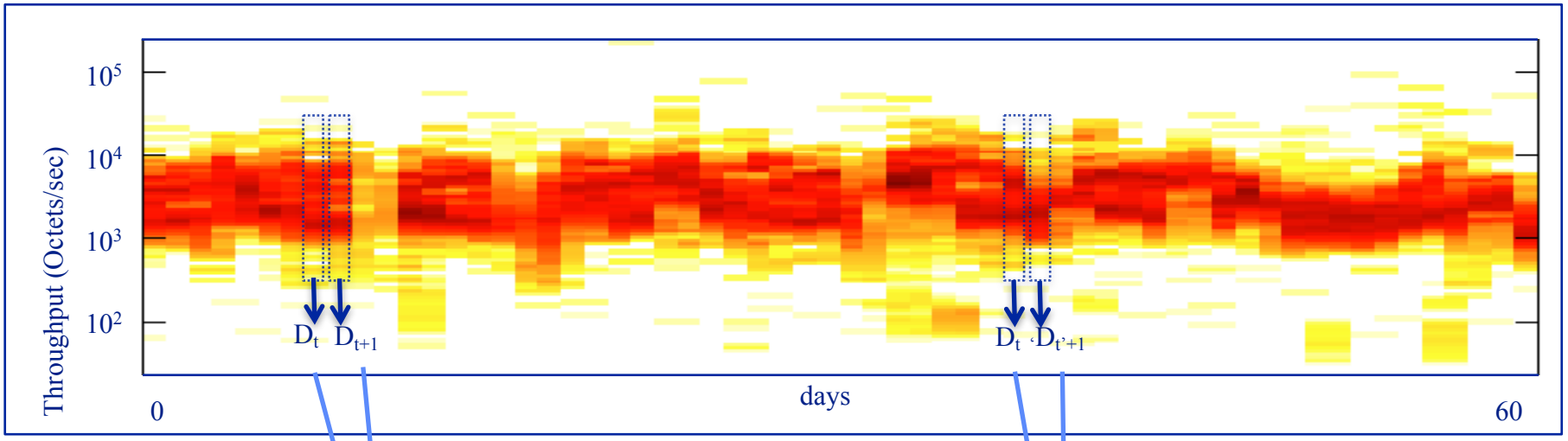
Input: streaming data

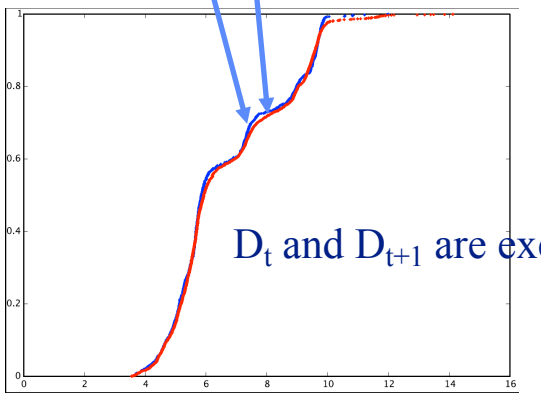


Output:

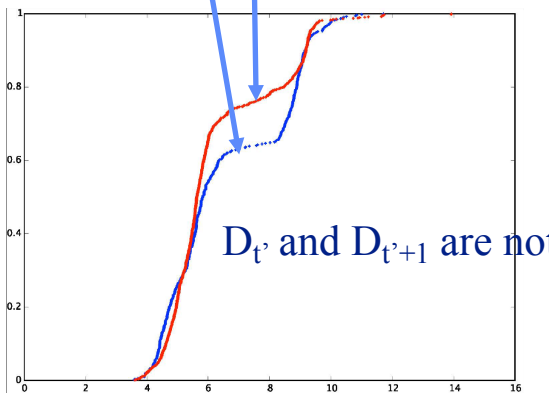
An example: Netflow data from ESnet

- **Checking exchangeable blocks by building cumulative histograms**





D_t and D_{t+1} are exchangeable



$D_{t'}$ and $D_{t'+1}$ are not exchangeable

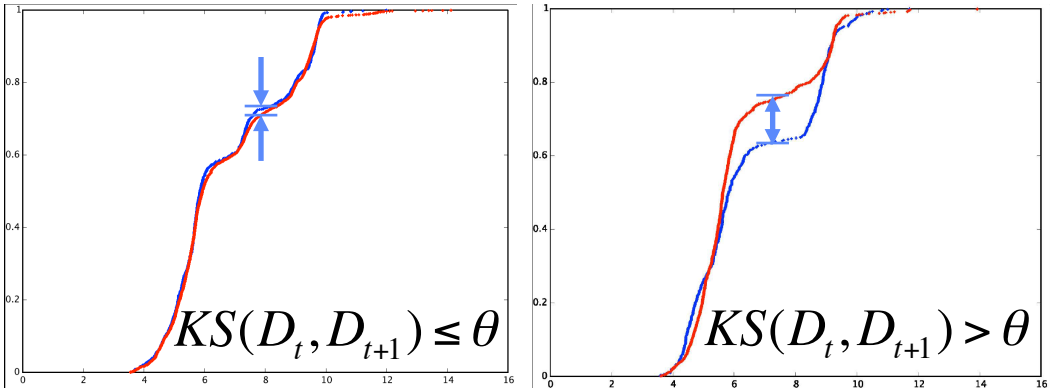
Kolmogorov-Smirnov test (KS test)

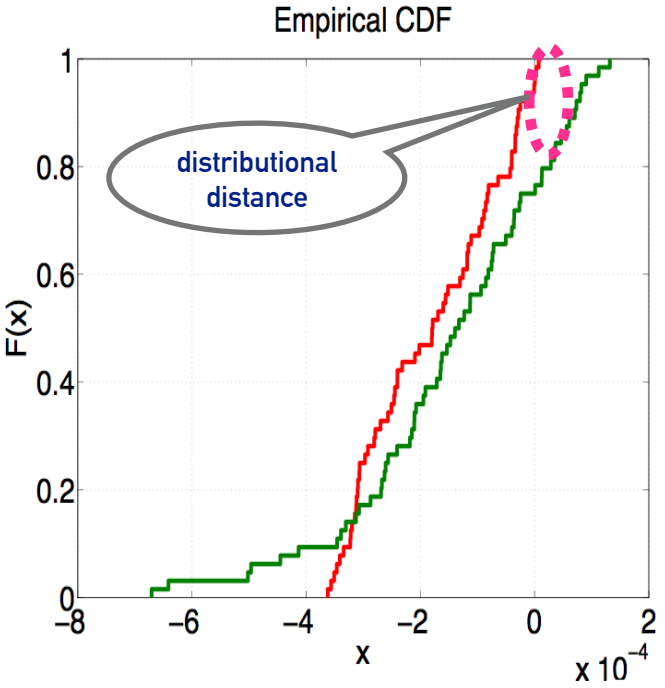
- **Statistical hypothesis testing by KS test to check exchangeable blocks**
  - **Measures distributional distance/similarity of two random variables**
    - $KS(D_t, D_{t+1}) = \max_l(|F_{D_t}(l) - F_{D_{t+1}}(l)|)$

KS score

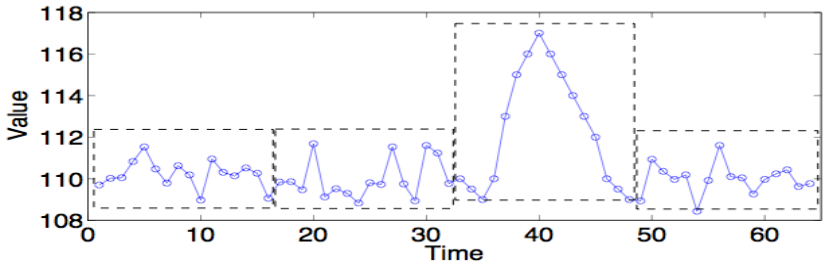
    - $F_D(l) = \frac{1}{N}\sum_{\substack{x_i \in X \text{ s.t. } 1\{x_i \le l\} \\ 1 \le i \le |D|}}$

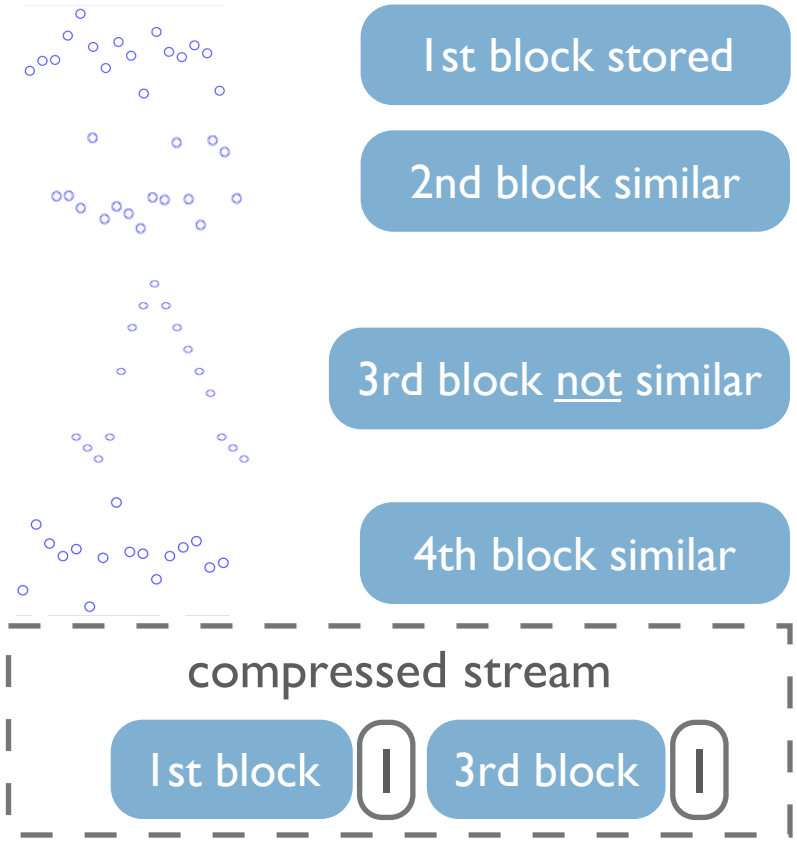
Empirical Cumulative Density Function (ECDF)



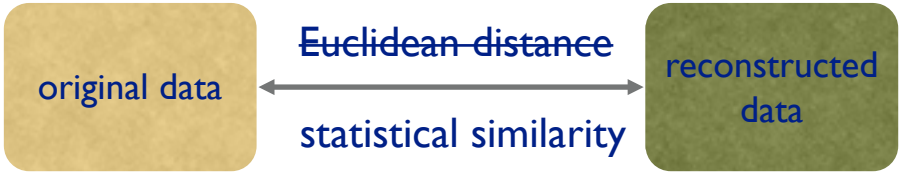


How IDEALEM works





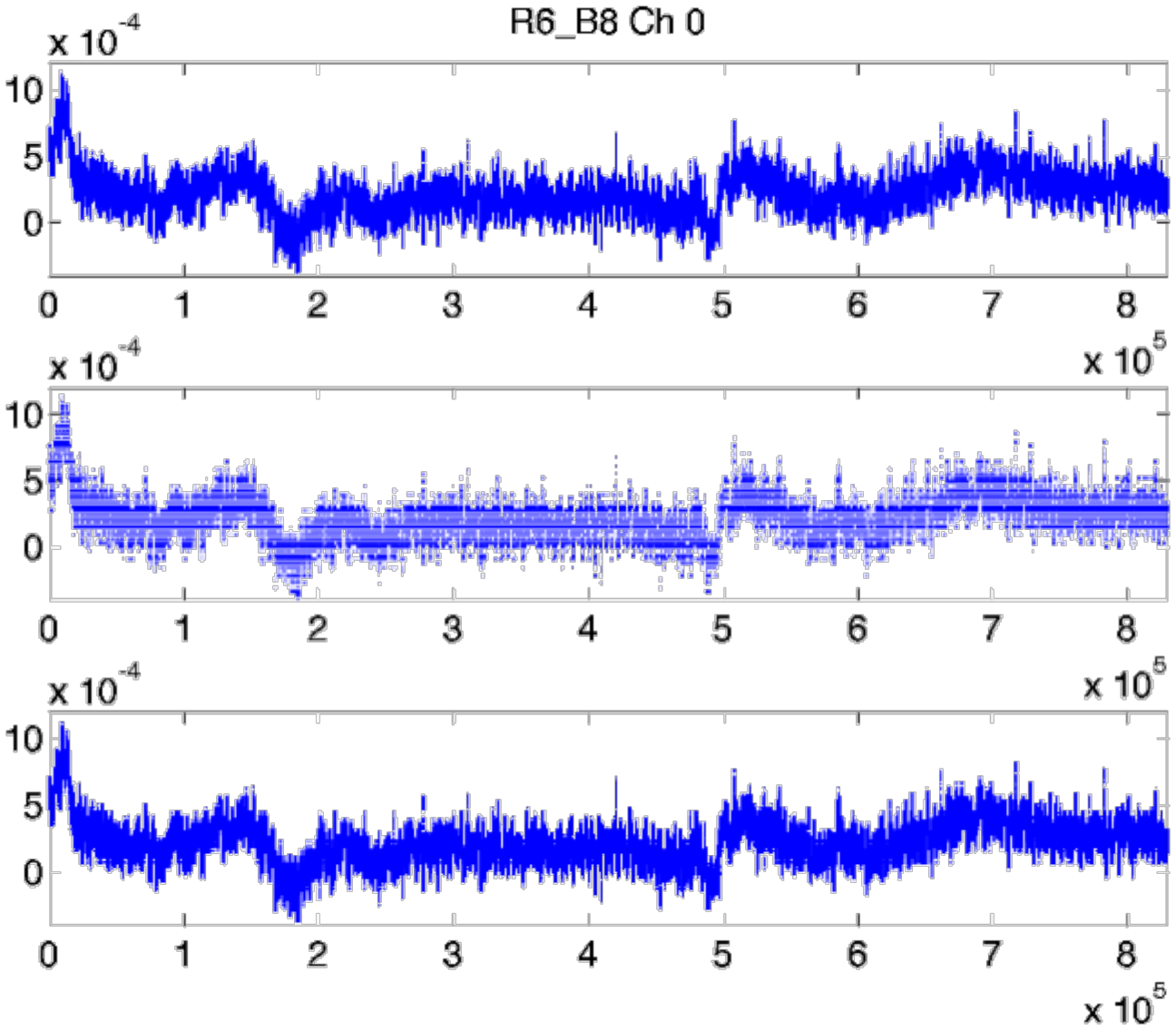
- **Breaks an incoming data stream into blocks of a fixed size**
- **Represents similar blocks with the one that appears earlier in the sequence**
- **Similarity here is based on statistical measure**
  - **Not on Euclidean distance**
  - **Kolmogorov-Smirnov test (KS test)**



Data Compression: Quick Review

- Two broad classes of data compression
- Lossless compression
  - gzip, 7-zip, PNG: work on repeated byte patterns
  - Floating-point values compression
    - FPC [Burtscher and Ratanaworabhan, 2009]: predictor+corrector
    - Difficult to compress because the lower order bits typically change
- Lossy compression
  - Common techniques: JPEG, MP3
  - Floating-point values compression techniques:
    - ISABELA [Lakshminarasimhan, et al, 2011]: sort + b-spline
    - Scalar Quantization Encoding [Iverson, et al, 2012]
    - zfp [Lindstrom 2014]
    - SZ [Di, et al, 2016]
- Challenges in compressing many scientific measurements:
  - Floating-point numbers are known to be hard to compress
  - "Random" fluctuations are hard to compress

IDEALEM Achieves CR>100



brain data (EEG) of a rat

original

state-of-the-art floating point compressor

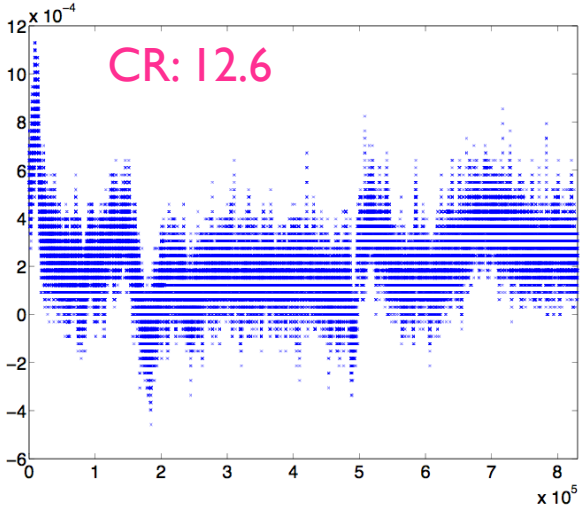
zfp -a 0.0004
CR: 12.6

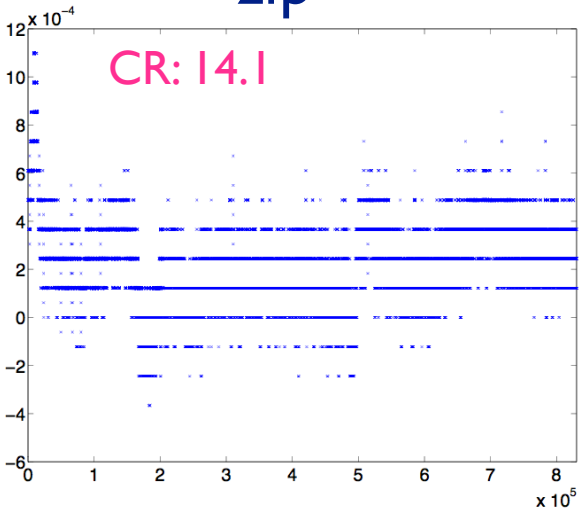
IDEALEM
CR: 106.6

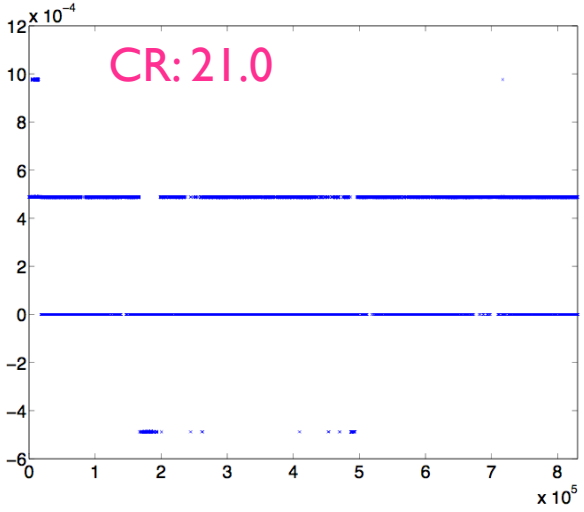
compression ratio (CR):
original size/compressed size

Compression ratio vs. Reconstruction Quality

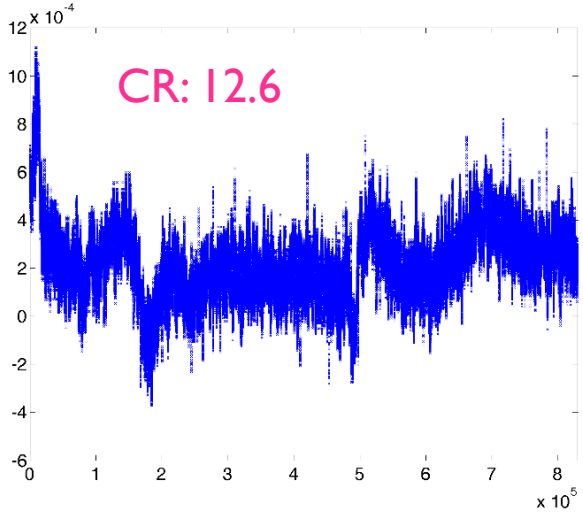
zfp

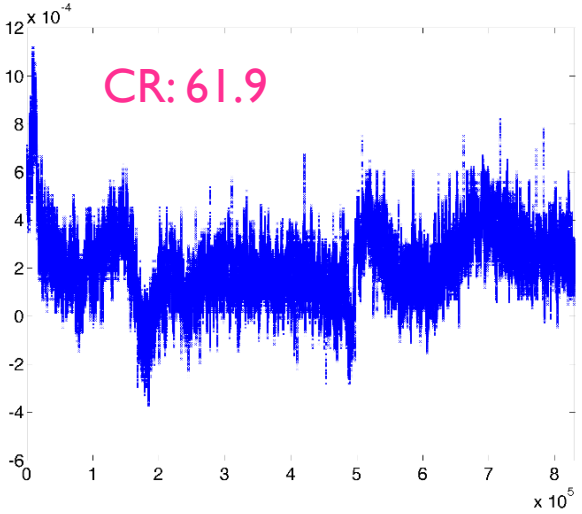


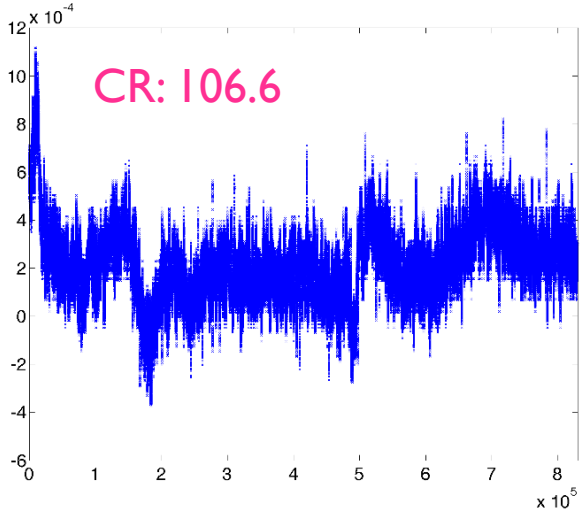




IDEALEM







An Application: µPMU for Monitoring Electric Power Grid

- Project µPMUs (present)
- Additional µPMUs (present)
- Additional µPMUs (prospective)

Berkeley LBNL
Alameda
PG&E
LBNL/CEC
LBNL/NSA

Riverside
SCE

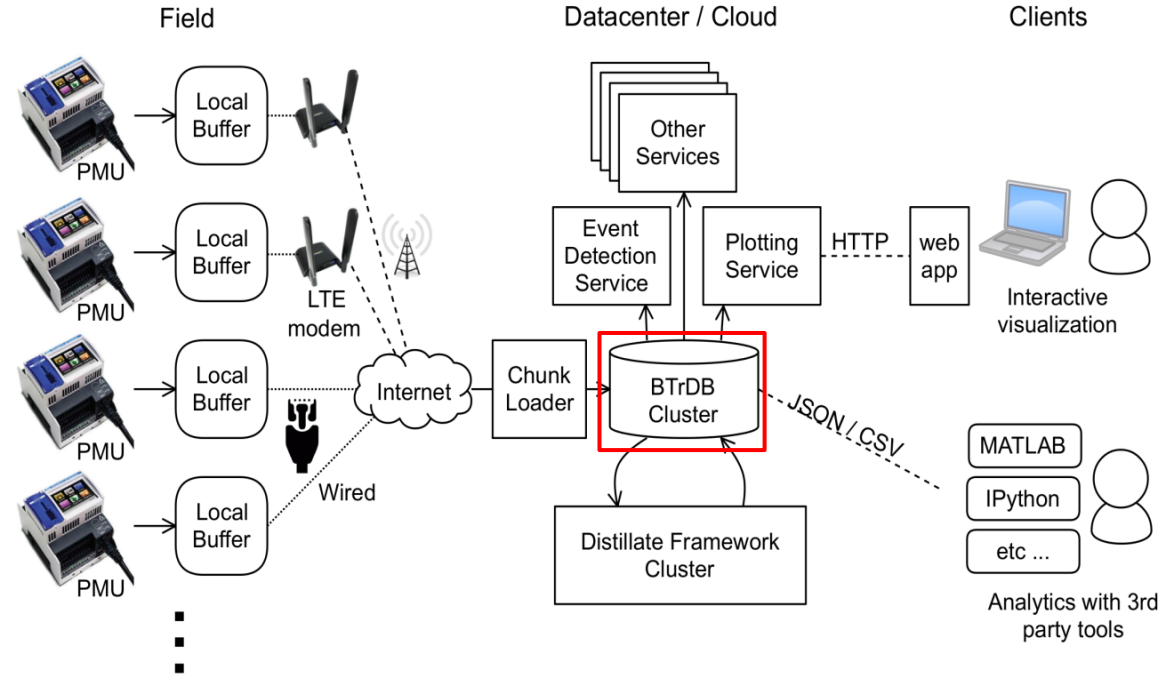
Sandia

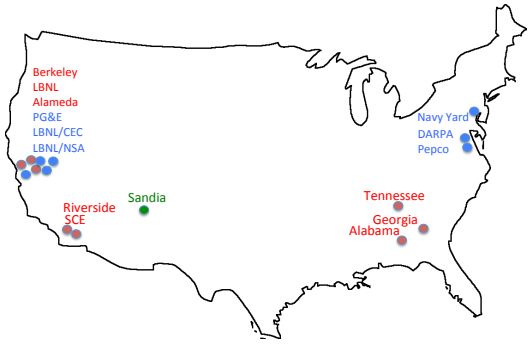
Navy Yard
DARPA
Pepco

Tennessee
Georgia
Alabama

12

Monitoring Electric Power Grid





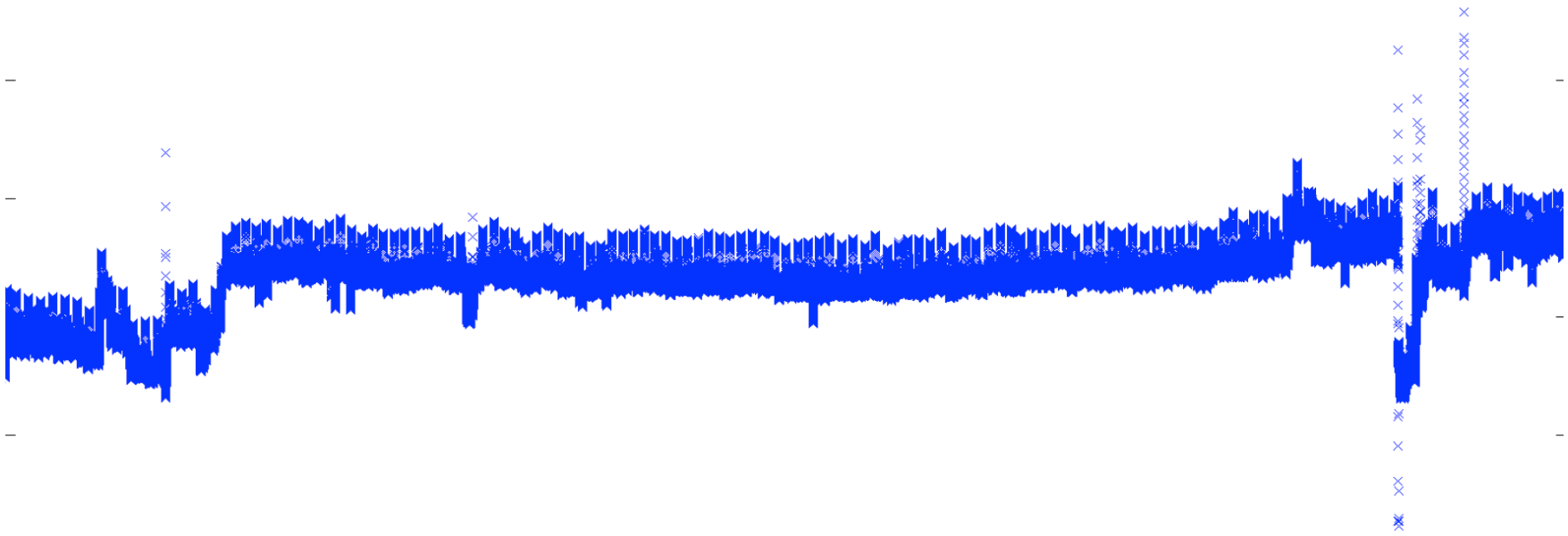
- Archiver / Database
- Stores (T, V) pairs
- Nanosecond precision
- Fault tolerant
- Highly scalable
- Unique abstraction
  - query range (ver)
  - insert values => ver
  - delete range => ver
  - query statistical (ver)
  - compute diff(v1, v2)

Challenges in µPMU Data

- **Data management challenges: Immense time series data distributed around the US**
  - Grid monitoring: 1,700 PMUs in North America generate 2M insertions per second (ips)
  - Grid usage data: 300M smart meters generate 0.33M ips
  - Analytics: 120M queries per second
  - Stream ALL the data to the cloud
- **Analytics challenges:**
  - Distillation infrastructure with extremely fast change set identification
  - On-the-Fly statistical summaries over a multi-resolution store
  - Multi-resolution search and process: e.g., find 'needle' events in immense haystacks instantly; drill down exponentially to analyze

Characteristics of µPMU Measurements

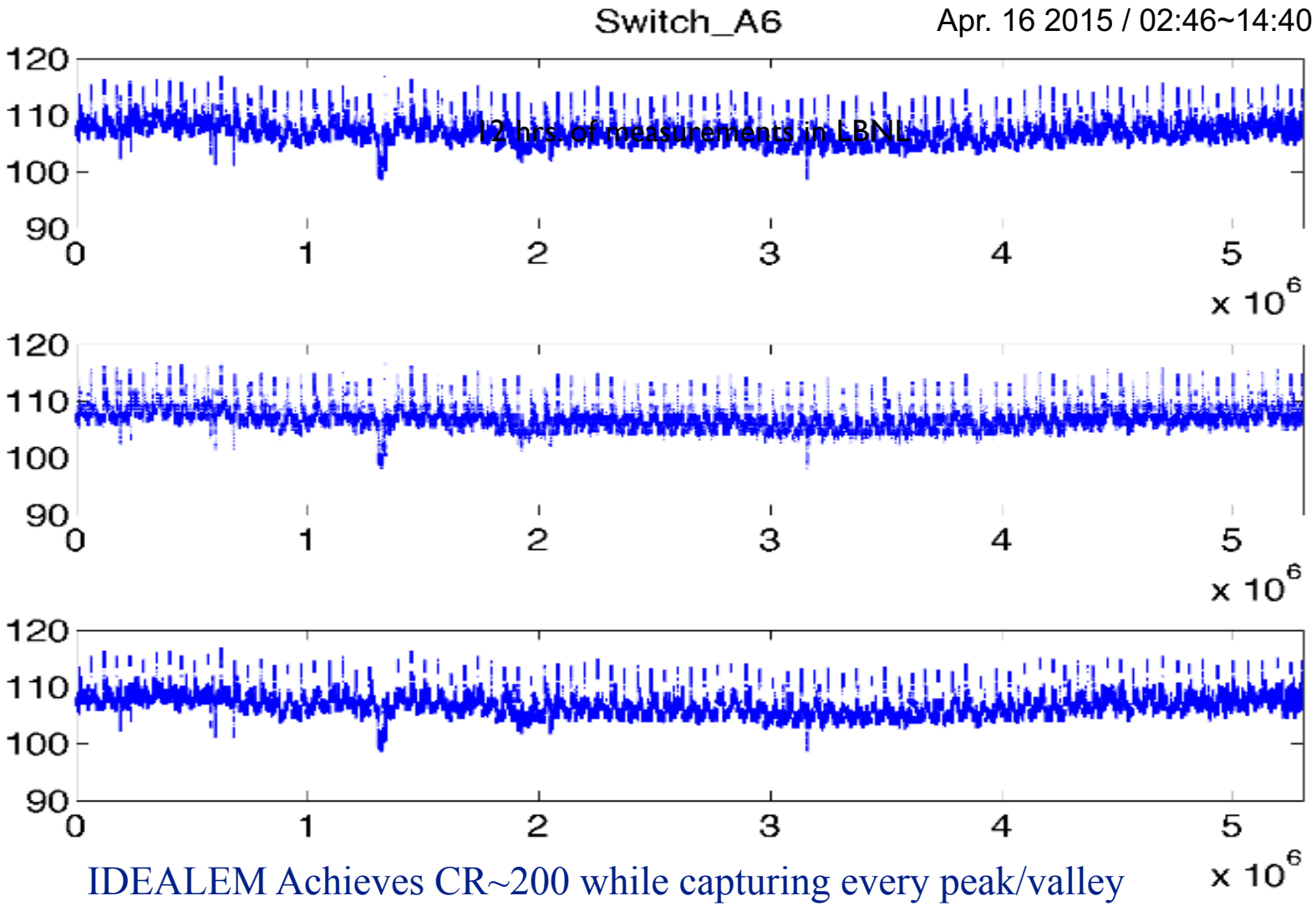
- **Numerical values: voltage, current, phase angles for voltage and currents**
- **Typically have a lot of "random" "small" fluctuations that are considered normal for the electric power grid system**
- **Occasionally, has relatively "large" changes that require attention or intervension**



What "Compression" Could Do?

- **Data compression is the science (and art) of representing information in a compact form**
  - Widely used in Internet, digital TV, mobile communication
- **For μPMU data,**
  - Compression will reduce the data volume to be sent around the data network
  - Compression will remove redundant information and make it easer to locate the interesting information
- **Previous compression approaches**
  - Top and Breneman (PES-GM 2013)
    - Lossless compression, CR around 2~3 (szip)
  - Gadde et al. (IEEE T. Smart Grid 2016)
    - Lossy compression (spatial and temporal redundancies), CR around 20
    - Feature for power system disturbance detection (NERC PRC 002)
- **IDEALEM for μPMU data**

IDEALEM for µPMU Measurements (1)



original

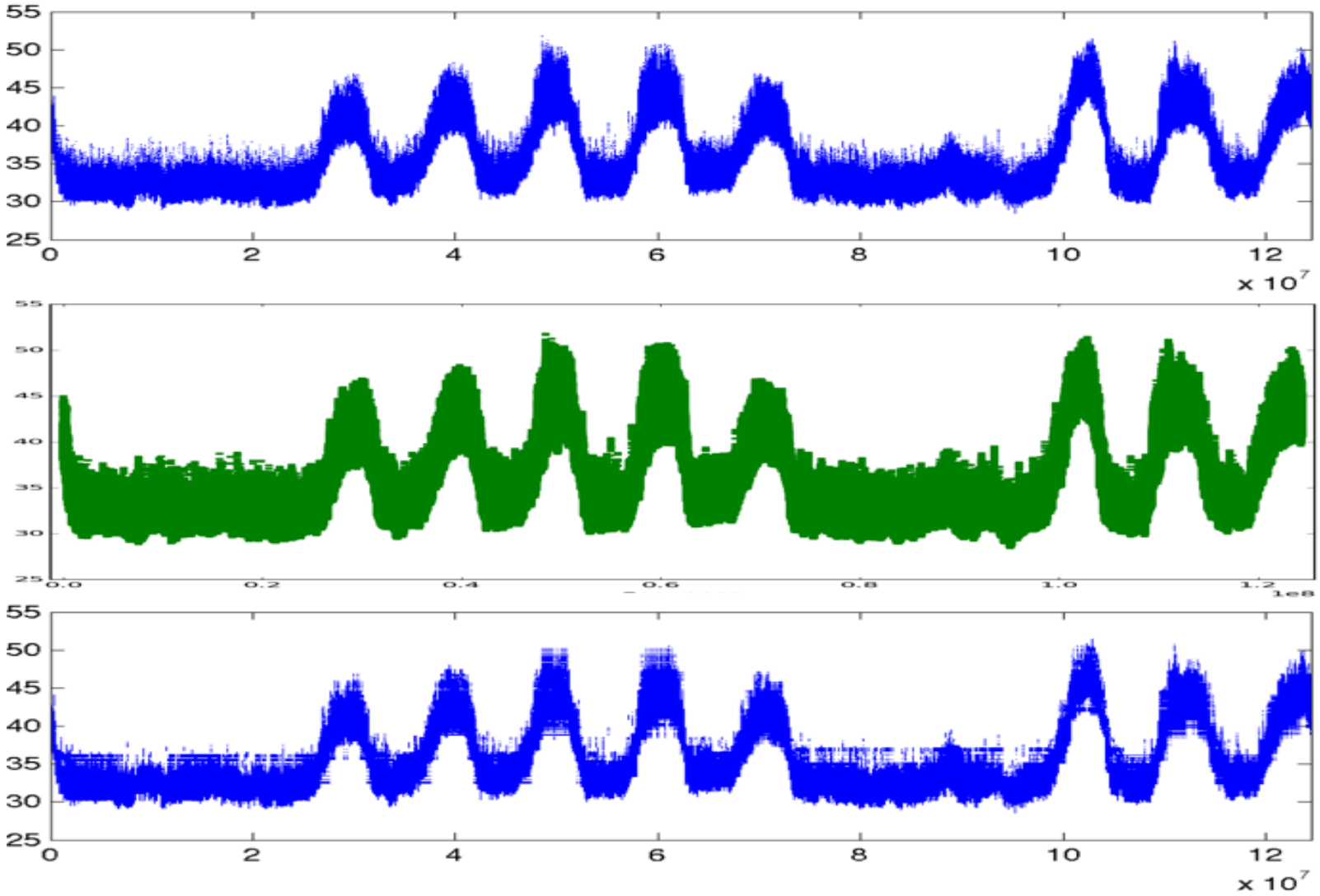
zfp –a 2
CR: 8

IDEALEM
CR: 189.3

IDEALEM Achieves CR~200 while capturing every peak/valley

IDEALEM for µPMU Measurements (2)

A6BUS1C1MAG (Apr. 18~Apr. 29, 2015)

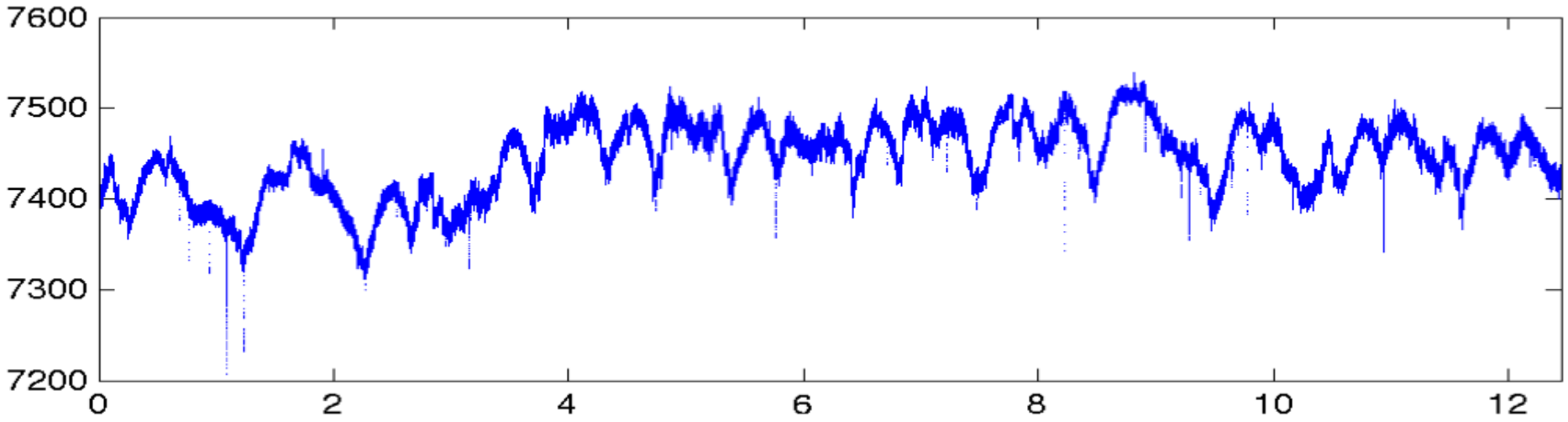


original

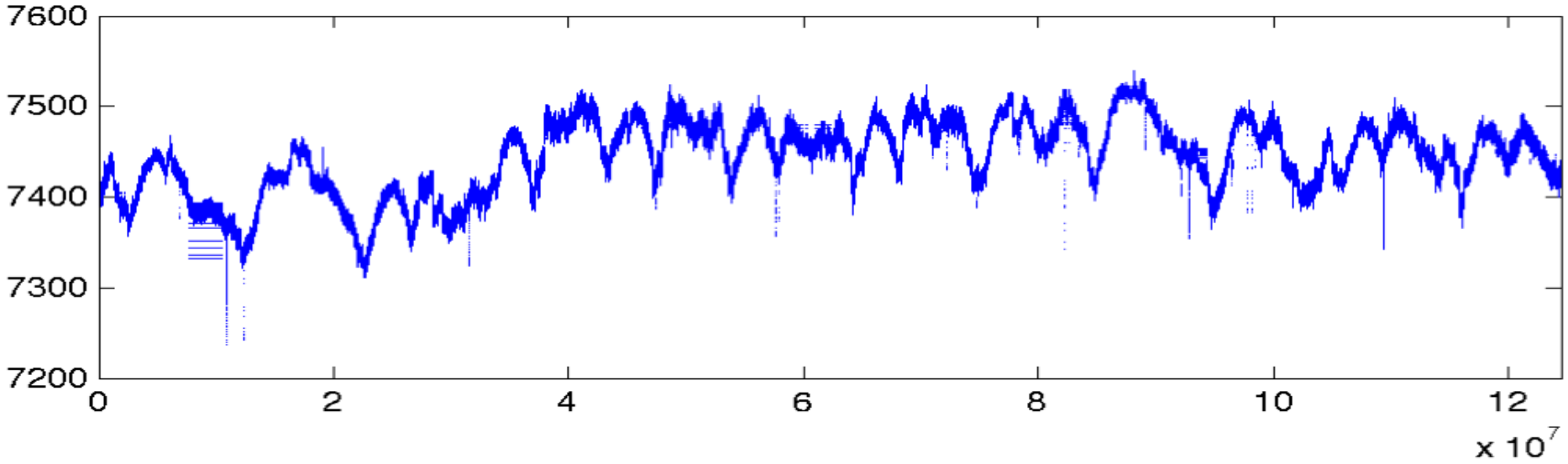
SZ
REL error bound 0.001
CR: 44.78

IDEALEM
CR: 242.3

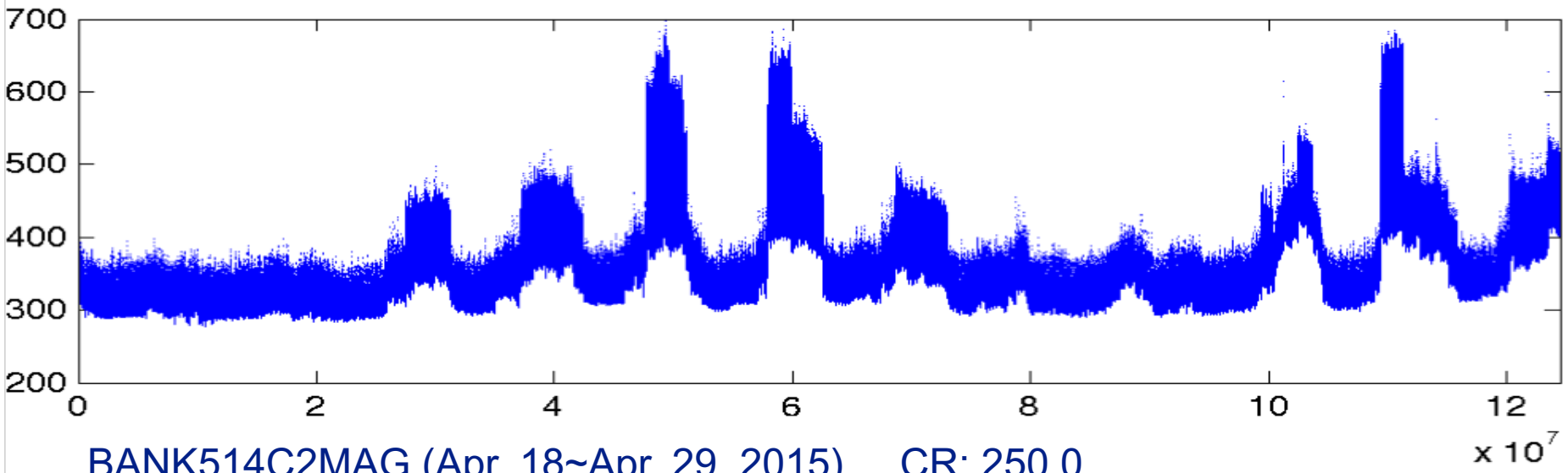
IDEALEM for µPMU Measurements (3)



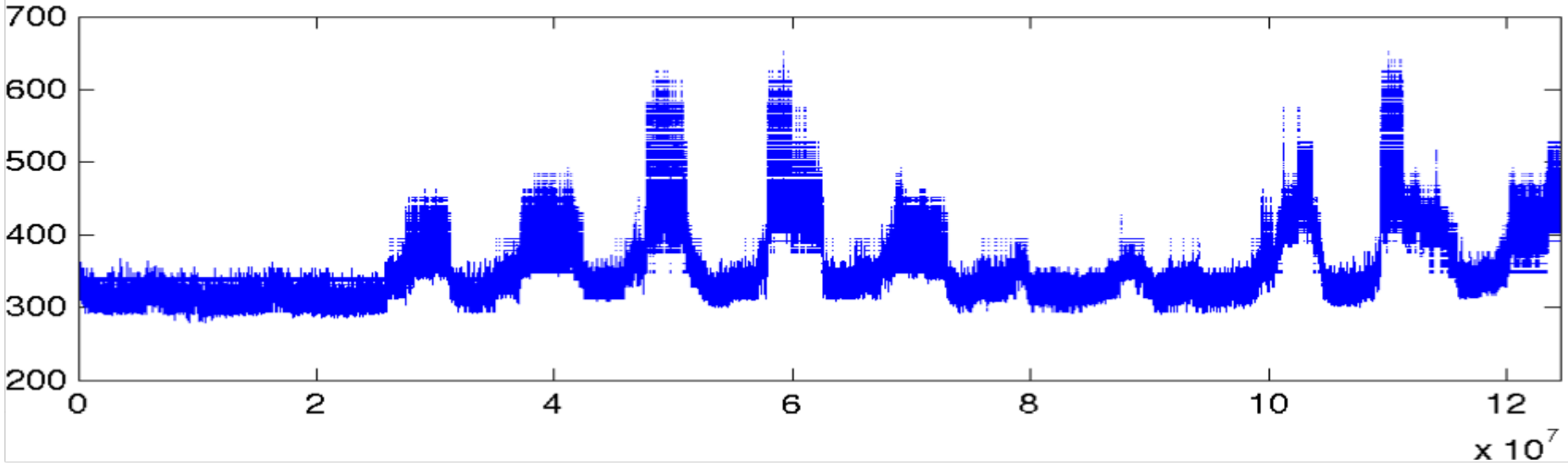
A6BUS1L1MAG (Apr. 18~Apr. 29, 2015) CR: 120.0



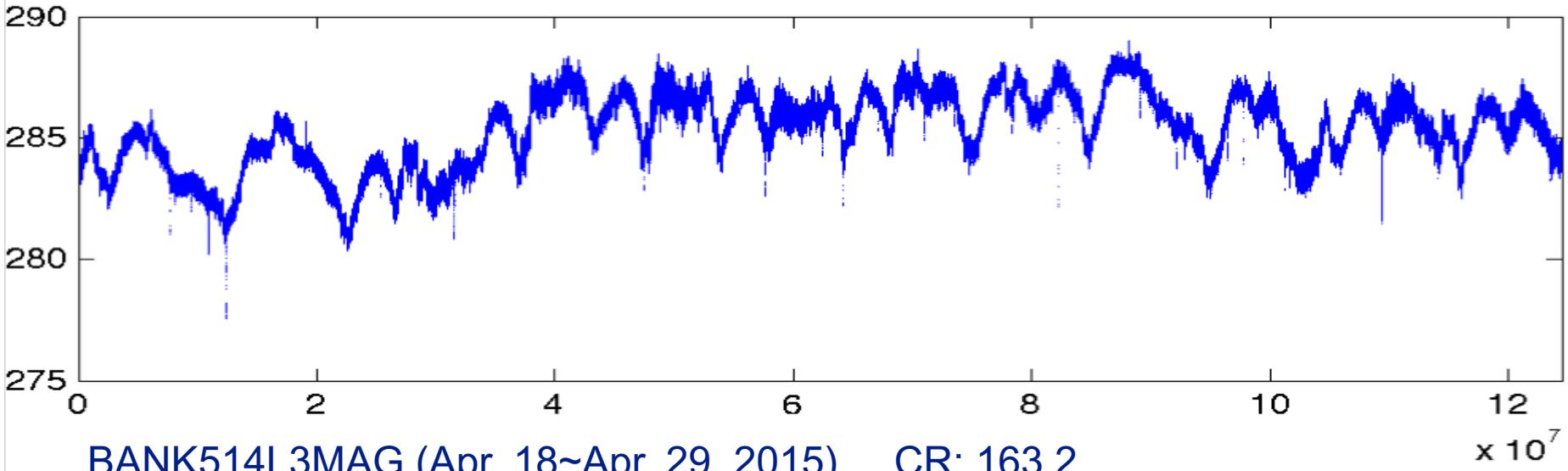
IDEALEM for µPMU Measurements (4)



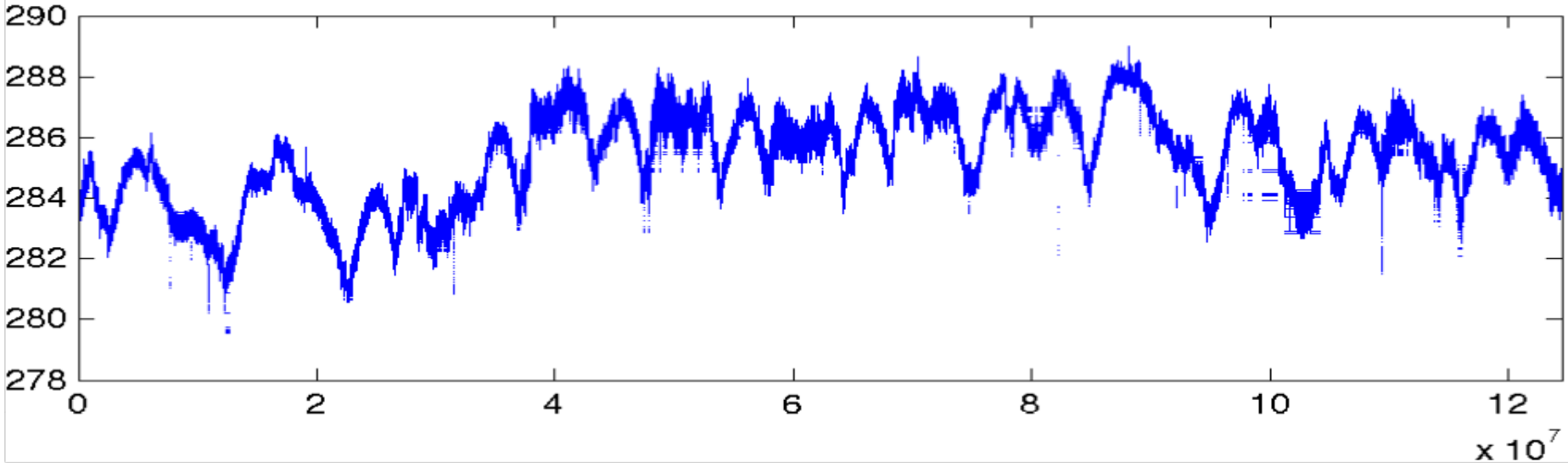
BANK514C2MAG (Apr. 18~Apr. 29, 2015) CR: 250.0



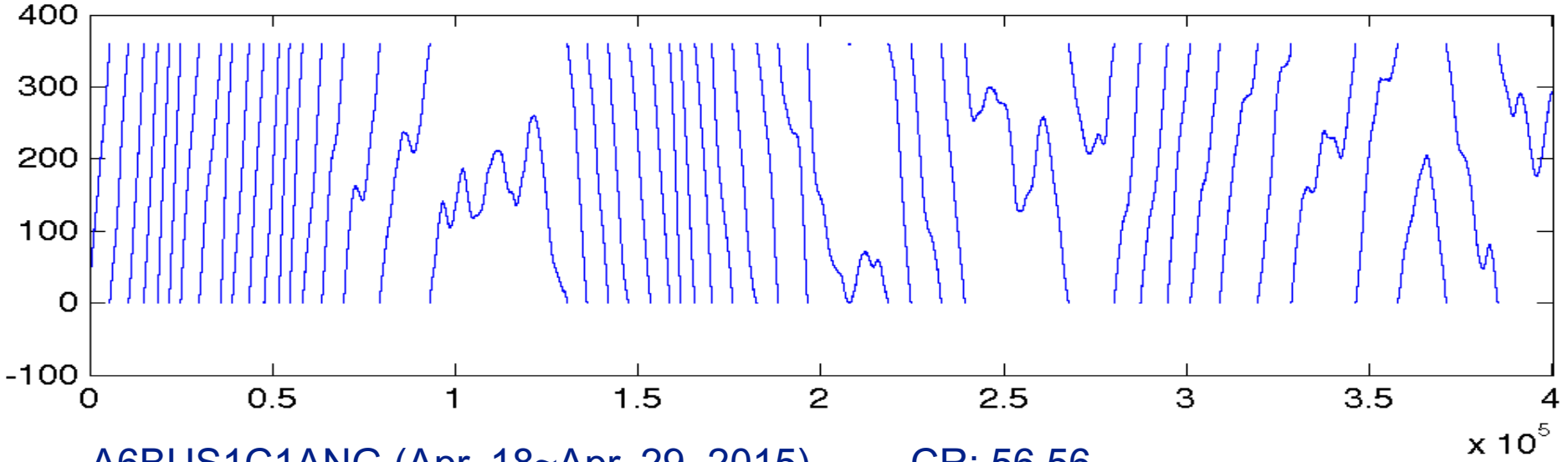
IDEALEM for µPMU Measurements (5)



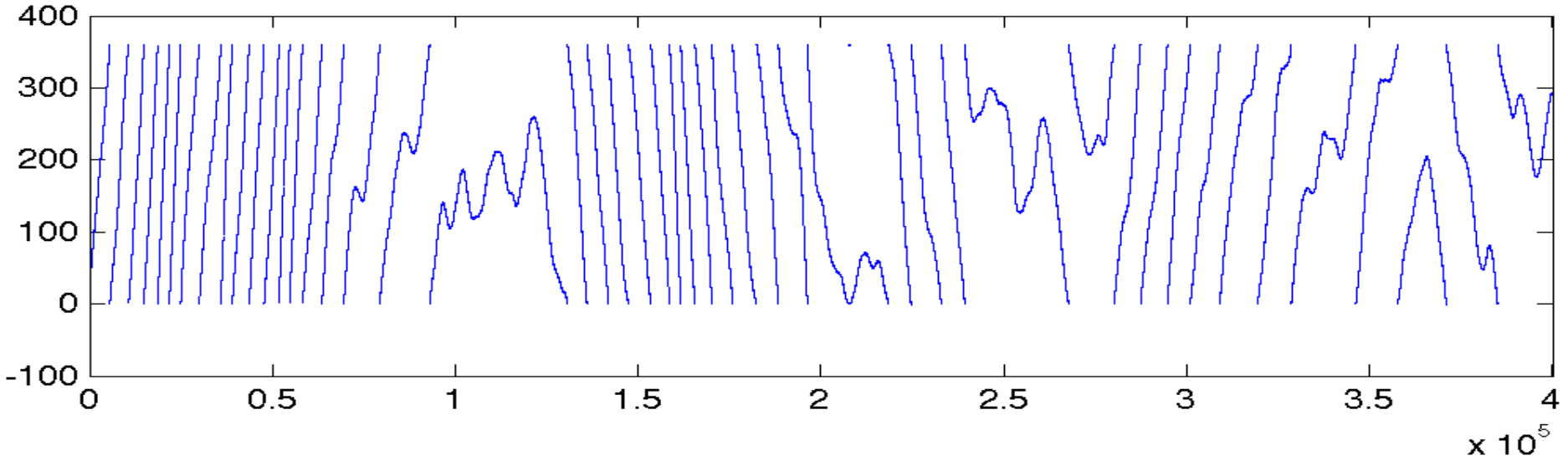
BANK514L3MAG (Apr. 18~Apr. 29, 2015) CR: 163.2



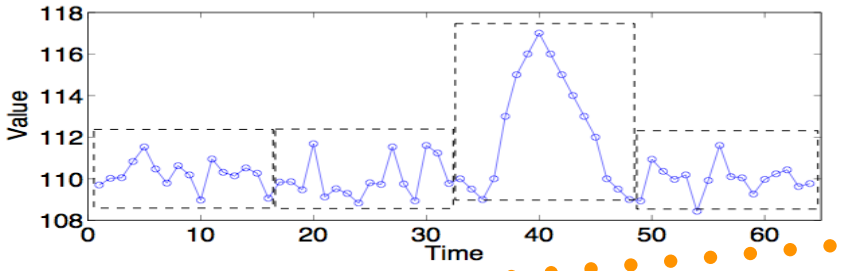
IDEALEM for µPMU Measurements (6) – Phase Angle Measurements



A6BUS1C1ANG (Apr. 18~Apr. 29, 2015) CR: 56.56



Three Key Parameters in IDEALEM



- **Block length**
  - how many samples?
- **Threshold for KS test**
  - how similar is similar?
- **Number of buffers**
  - how many buffers?

1st block stored

2nd block similar

3rd block not similar

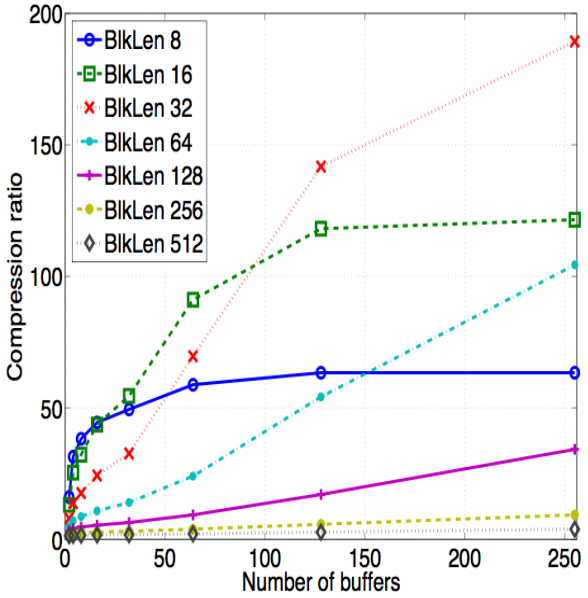
4th block similar

compressed stream

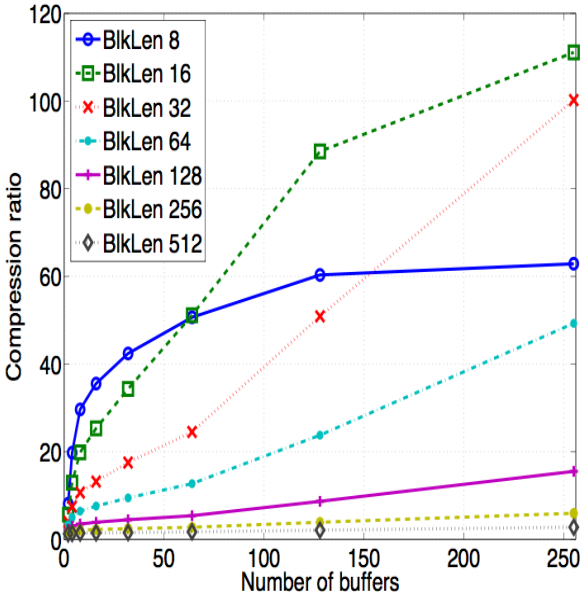
1st block | 3rd block |

How Three Key Parameters Affect Compression Ratio

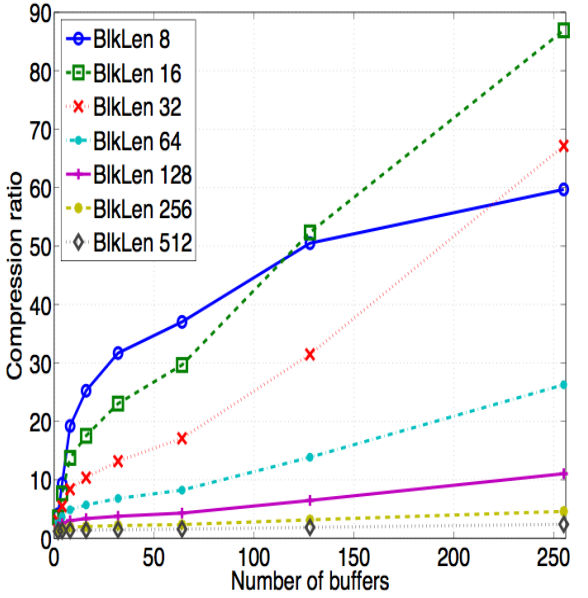
power grid monitoring data



threshold: 0.01



threshold: 0.05

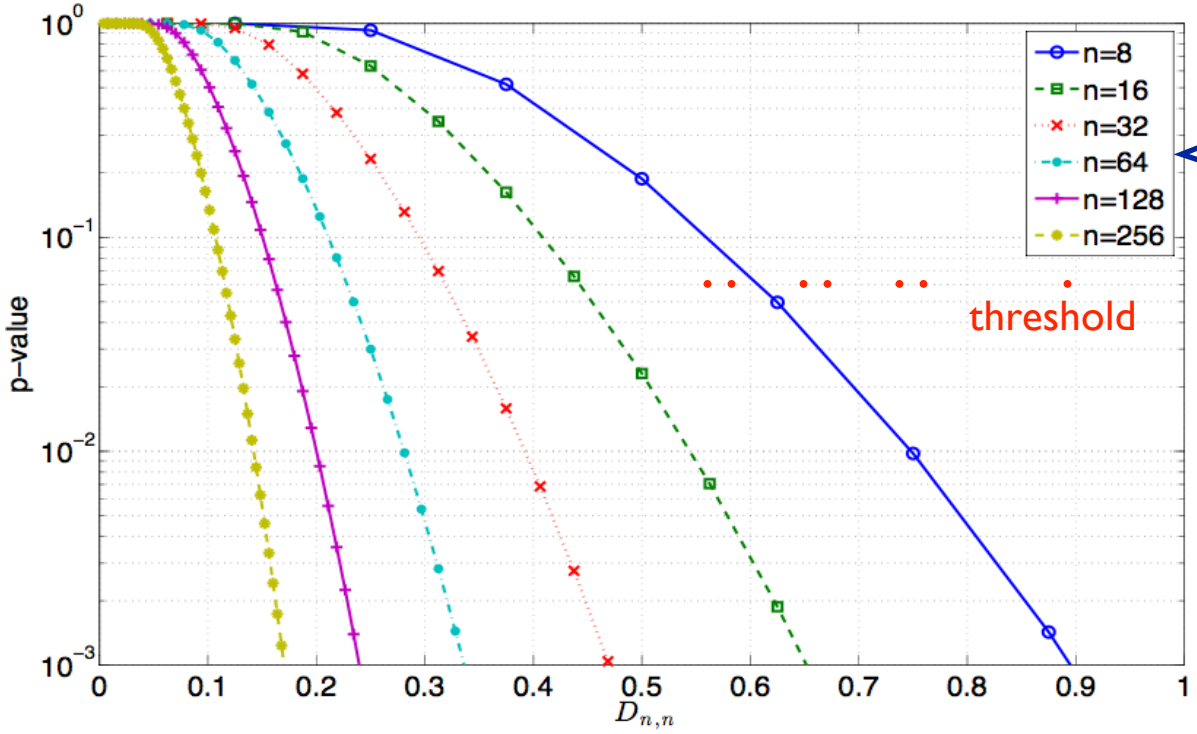


threshold: 0.1

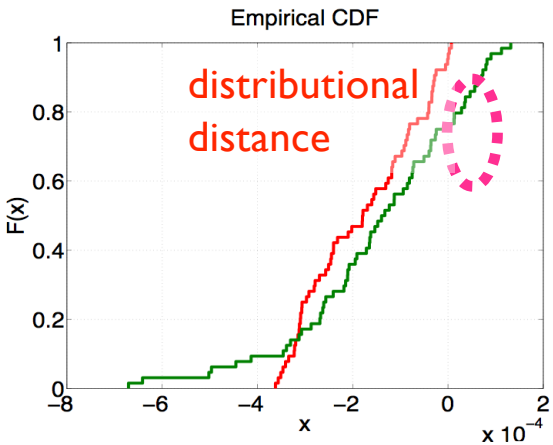
- **Two parameters on compression ratio (CR)**
  - **CR ↑ with threshold for KS test ↓**
  - **CR ↑ with number of buffers ↑**
- **Effect of block length (BlkLen) is not immediately apparent**
- **Small memory usage: 128KB for BlkLen=32 and 255 buffers**

Limits on Achievable Compression Ratio

- **Given a block length n, the maximum achievable CR of IDEALEM encoder with multiple buffers is $8 \cdot n$**
  - **assuming double precision floating-point format (8 bytes)**
- **Large BlkLen n potentially increases CR, but it also increases difficulty of passing the KS test**



large n makes it difficult to pass KS test for the same distributional distance



More application areas

- **Statistical analysis enables estimating future events in various applications. For example,**
  - Financial market analysis
  - Environmental study (e.g. extreme weather, climate change)
  - Energy usage analysis
  - Social network media analysis
  - Traffic analysis
  - System performance monitoring analysis
- **IDEALEM**
  - Enables efficient data reduction on the large streaming data
  - Provides accurate statistical analysis without loosing the underlying data distribution
  - Can also be applicable to large data archives (offline data)

Summary

- **IDEALEM is a new class of compression methods**
  - measures distance based on statistical similarity
  - not traditional Euclidean distance (L^2-norm)
- **IDEALEM can reduce data volume by more than 100-fold, while retaining key features from original data**
  - Application to large, high frequency streaming data as well as large offline data archives
- **Fast enough execution time and small memory footprints to be used on resource limited devices for real time compression**

More information

- SC'16 demo info including IDEALEM iOS 10 demo app
  - http://sdm.lbl.gov/asim/idealem.html
- Software downloads
  - Available for commercial and non-commercial use
  - http://datagrid.lbl.gov/idealem
- License info
  - http://ipo.lbl.gov/lbnl2013-133/
  - U.S. Patent pending, serial no. 14/555,365
- Email SDMSupport@LBL.Gov
- SDM Group http://sdm.lbl.gov
- LBNL http://www.lbl.gov