Dealing with BIG data:

Data quality assessment and visual analytics

Sirish L. Shah
Department of Chemical and Materials Engineering
University of Alberta, Canada

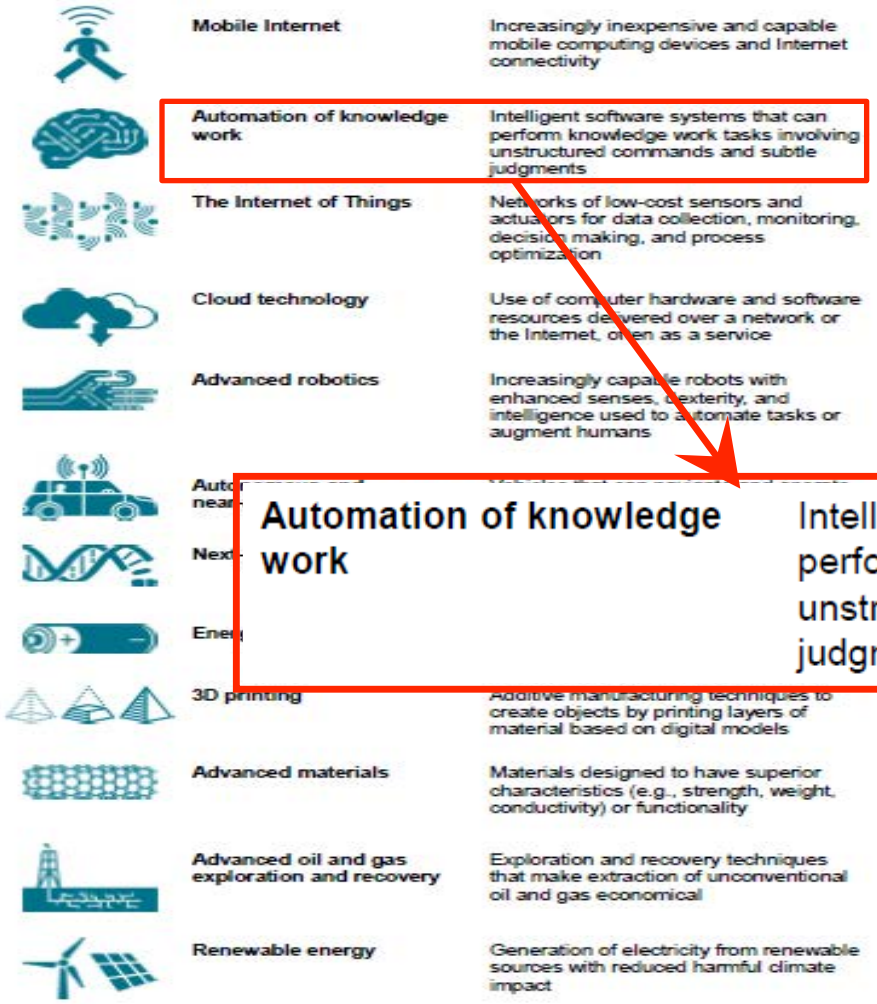
Information fusion and the digital tsunami [2]

Main message

- BIG Data will open incredible opportunities
- Will require skills in statistical learning, signal processing, image processing, data base design/ organization/sorting, machine learning and methods in computing science.

"Data constitutes a new natural resource, which promises to be for the 21st century what steam power was for the 18th, electricity for the 19th and hydrocarbon for the 20th."
V. Rometty, IBM CEO

Disruptive technologies



Report by McKinsey Global Institute:

Disruptive technologies: Advances that will transform life, business, and the global economy

Automation of knowledge work

Intelligent software systems that can perform knowledge work tasks involving unstructured commands and subtle judgments

"By 2018, the US could face a shortage of up to 190,000 workers with analytical skills"

McKinsey Global Institute

"The sexy job in the next 10 years will be ~~statisticians~~." *Data Scientists?*

Hal Varian, Prof. Emeritus UC Berkeley
Chief Economist, Google

Hal Varian Explains...

The ability to take **data** – to be able to **understand** it, to **process** it, to **extract value** from it, to **visualize** it, to **communicate** it's going to be a hugely important skill in the next decades, not only at the professional level but even at the educational level for elementary school kids, for high school kids, for college kids. Because now we really do have essentially free and **ubiquitous data**." – Hal Varian

BIG data, a growing torrent

- > 6 billion mobile phones in place in 2016
- 30 billion pieces of content shared on facebook every month
- 40% projected growth in global data/year
- $300 billion annual value to US health care
- $600 billion value to retailers
- 190,000 data analytics positions in US
- 1.5 million data-literate managers needed in the US

The Challenge of BIG data

- Are we all truly Data-Literate?
- Data ➡ Information ➡**ACTION**
- Need to have required skills in 'Informatics' for all engineers and scientists.
- The world is becoming data-centric.
- BIG Data comes in many 'shapes', sizes, and 'colours'.
- Data-based decision is better than intuition.

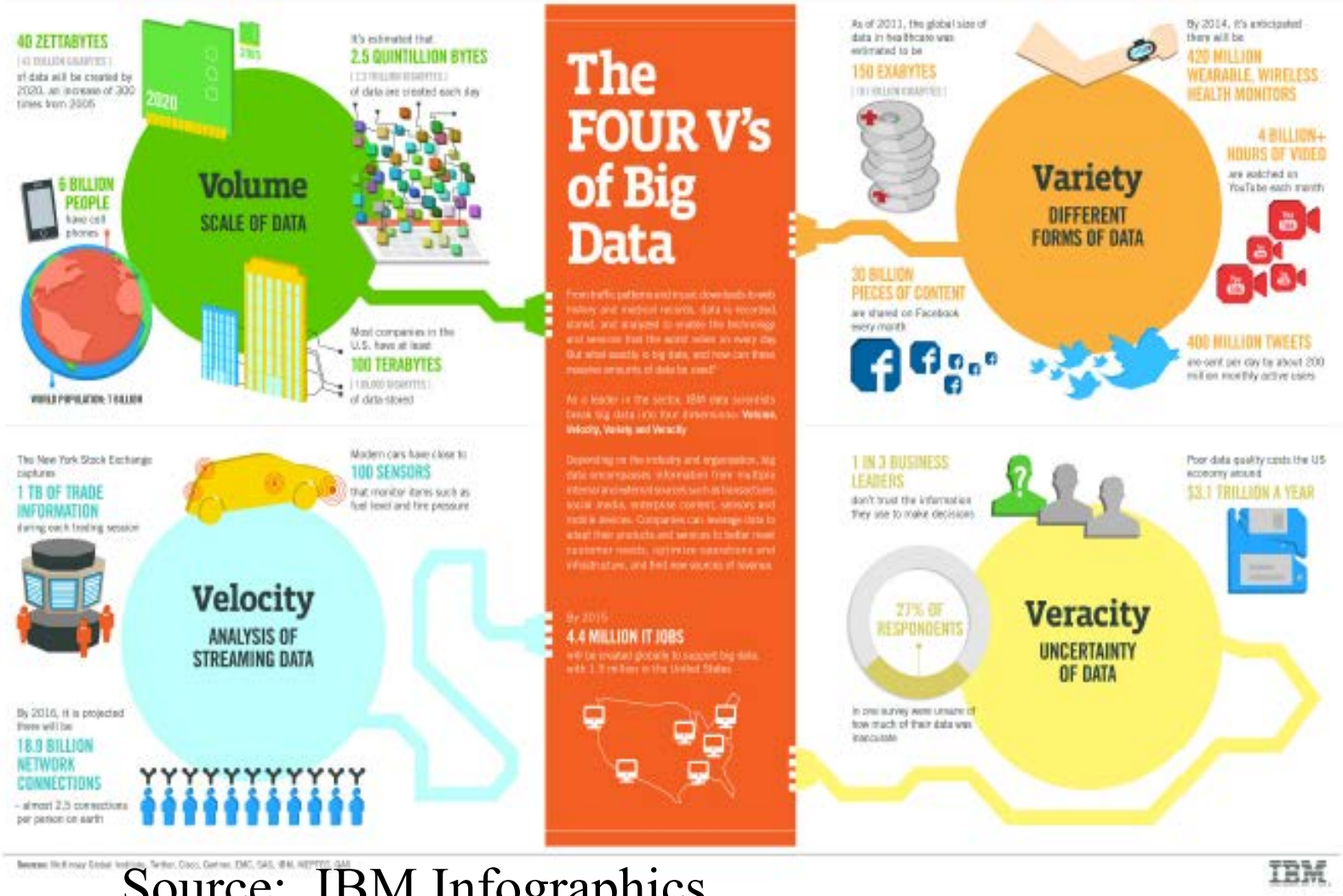
Data, data everywhere….

- What is BIG data?
- 'BIG' is relative. But generally terrabytes, petabytes and more…but scale will change
- Many publicly available reports:
  - IBM "BIG" data infographics + reports
  - McKinsey Global Institute report

Dealing with BIG Data…

- Look at integrated solutions that give a holistic picture of the process.
- Challenge of big data:
  - More data … less Insight;
  - V^4: Volume, Velocity, Variety, Veracity,…

V⁴ of Big Data

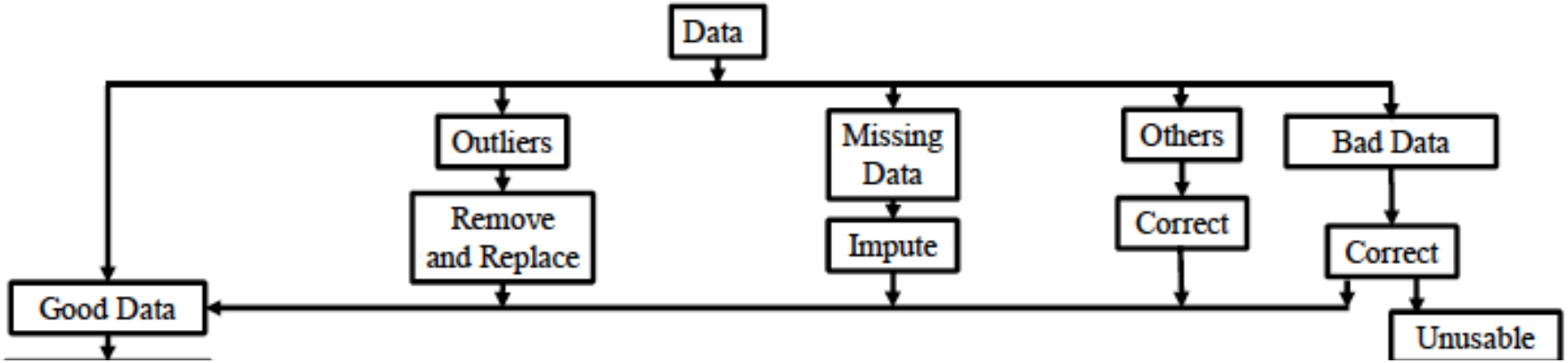


Source: IBM Infographics

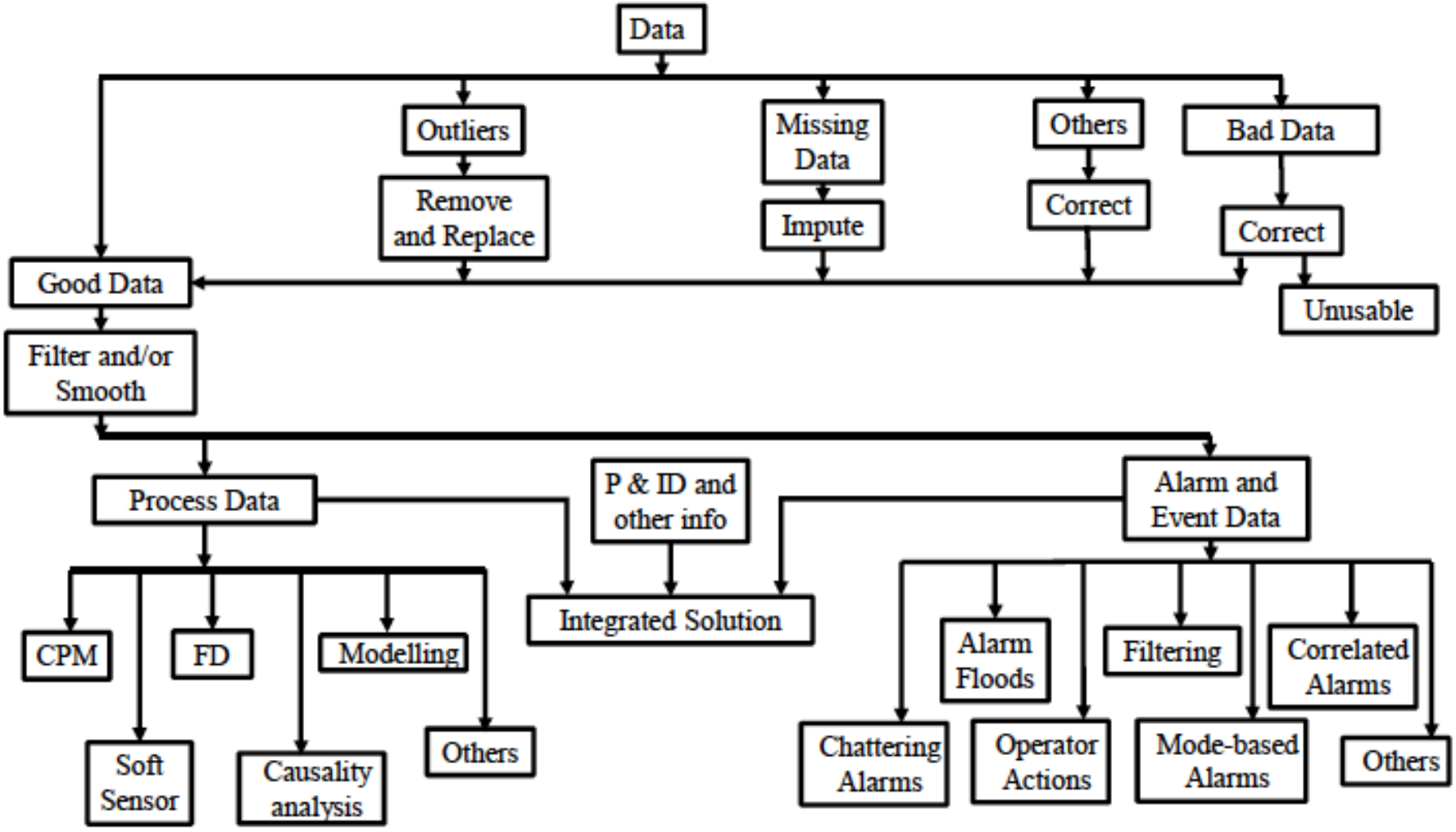
Dealing with BIG Data

- Look at integrated solutions that give a holistic picture of the plant.
- Challenge of big data:
  - More data … less Insight;
  - V^4: **V**olume, **V**elocity, **V**ariety, **V**eracity,…
- What about **V**alue (for data quality)?
  - Archives and compression
  - Bandwidth limitations
  - Data segmentation and visualization tools

Challenges in dealing with industrial data



Challenges in dealing with industrial data



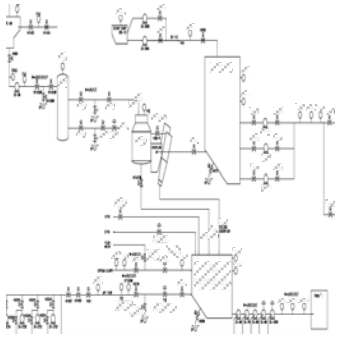
Ingredients of smart process data analytics

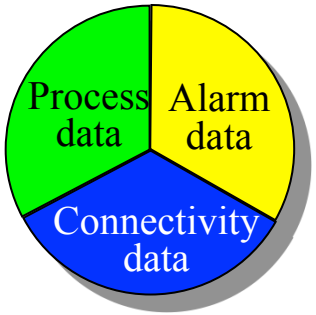
2008-11-06 14:40:01	3644.24	3444.58
2008-11-06 14:40:02	3641.41	3441.60
2008-11-06 14:40:03	3631.96	3446.31
2008-11-06 14:40:04	3626.36	3452.57
2008-11-06 14:40:05	3627.51	3454.23
2008-11-06 14:40:06	3618.98	3465.41

Process Data (DA)
- very simple and easily available
- not fully utilized for alarm design

2005-04-15 21:33:25	MES	BAD
2005-04-15 21:33:40	MES	LO
2005-04-15 21:33:55	MES	HIHI
2005-04-15 21:34:02	SETPOINT	
2005-04-15 21:34:41	MES	HI
2005-04-15 22:37:15	RESULT	CI

Alarm Data (A&E)
- more complicated than process data
- main resource for Alarm Management software

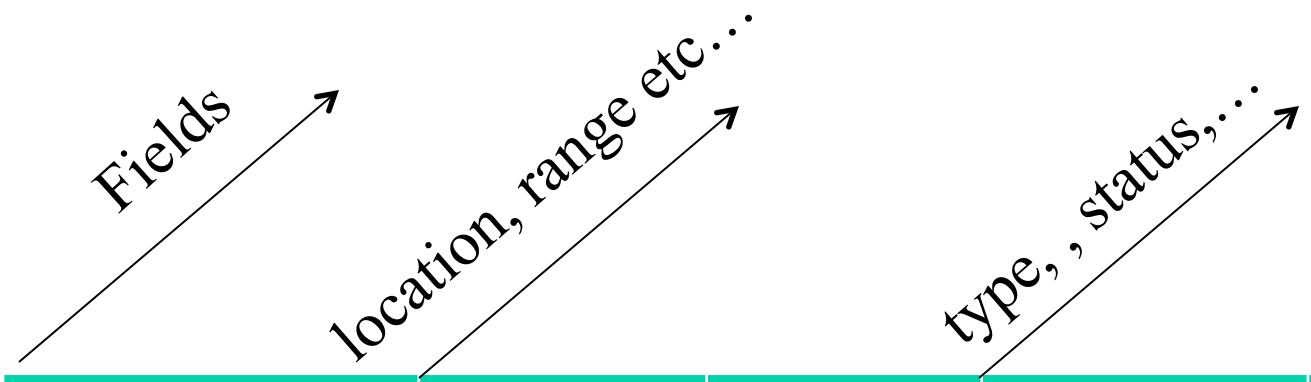




What about event logs
and workflow mining & conformance ?

Connectivity Data (P&ID)
- more complicated than process and alarm data
- main resource for alarm rationalization

Process industry data streams



Time	Sensor 1	2	Actuator 1	…….2
03/03/14 00.01				
00.02				
….				
….				
….				

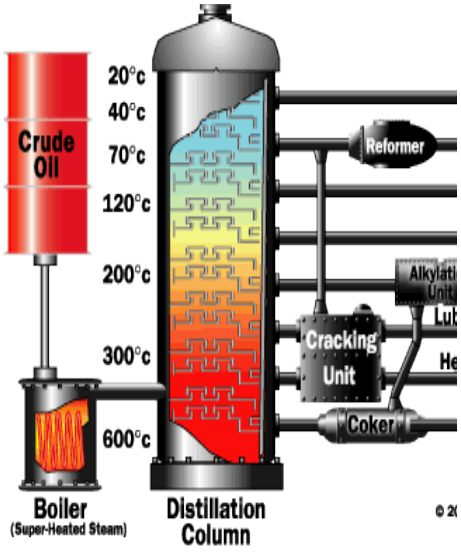
Assessment of data quality

- Routine industrial process **data** is **plentiful**.
- However, its **usefulness** for **process and performance monitoring** may be questionable.
- Thus, there is a need to consider how to determine which **segments of the data set** are informative enough for their intended end use.

Data quality

- Data collection/archiving is basic infrastructure.
  - Correct compression tools should be used.
  - Ensure that reconstructed data is informative.
- Obtain measures of data quality or segment/ partition data and use segments that have the necessary information.

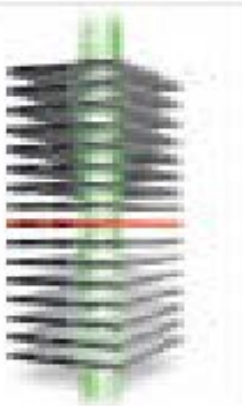
Data Compression in the Process Industry



X_{raw}

Compression

X_{comp}



Data Historian

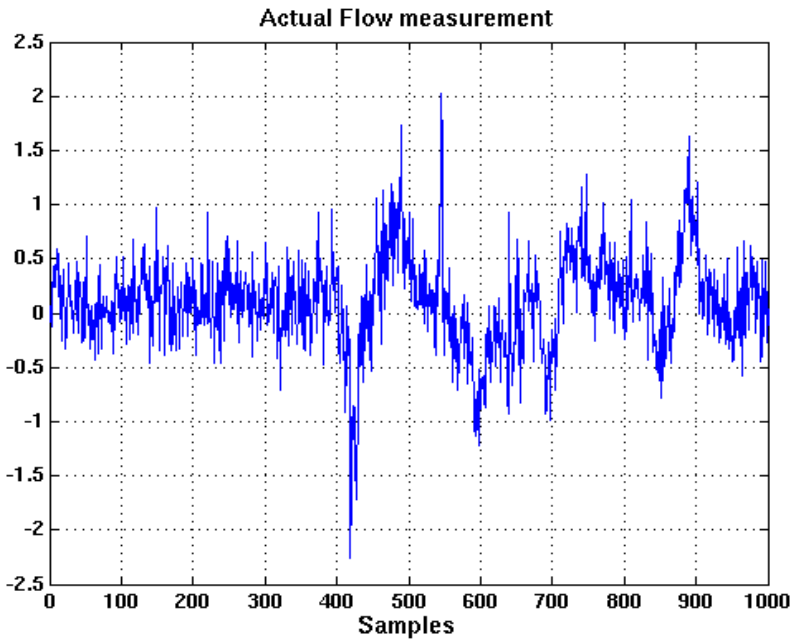
Decompression

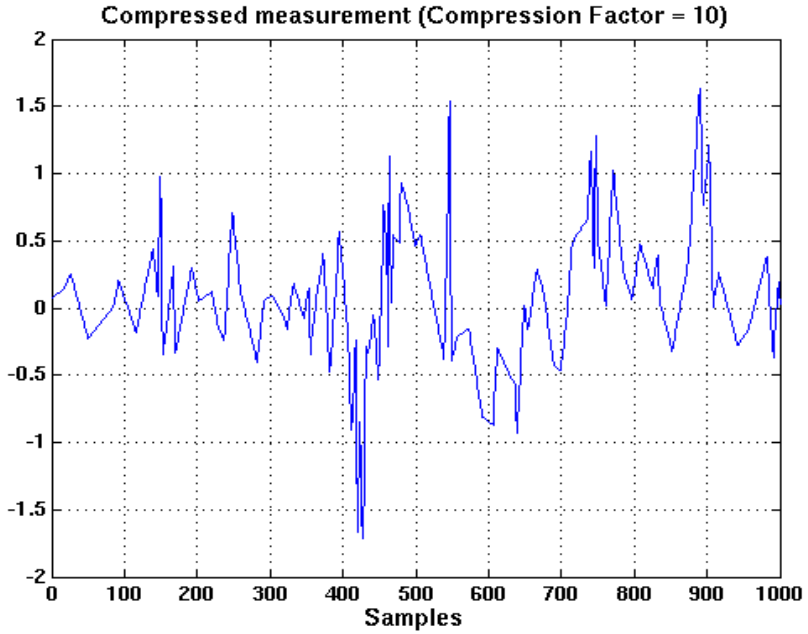
Data Analyst

$X_{reconst}$

$X_{reconst} \neq X_{raw}$

Data Compression



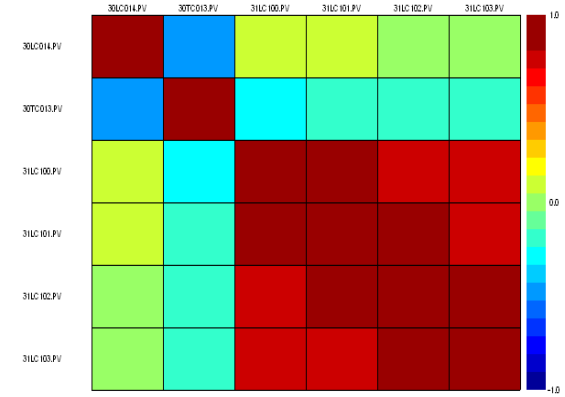


- Minimizes data stored in archive.
- Modifies correlation structure & relationships.
- Affects every subsequent analysis of the data.

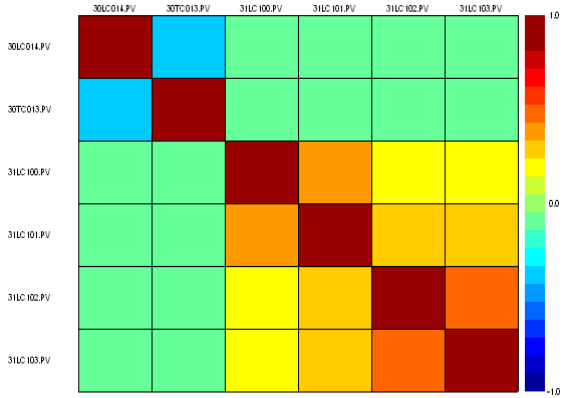
CF = N / (N - d);
N: Total no. of samples d: No. of zero second derivatives of the samples

Effect of Compression on Correlation Structure Refinery Process Data

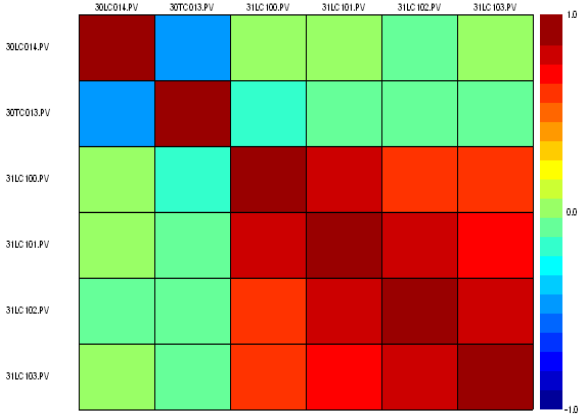
Uncompressed Data Correlation Color Plot



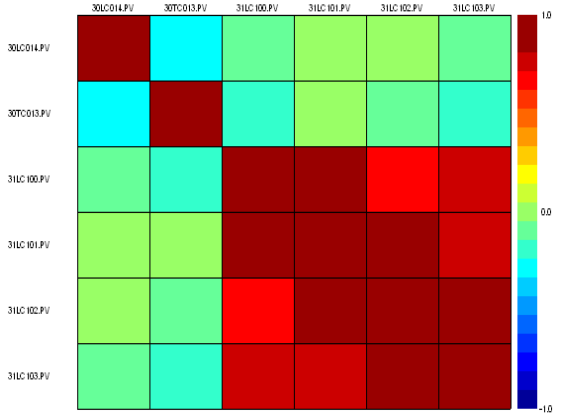
Correlation Color Plot of Reconstructed Data from Swinging Door algorithm



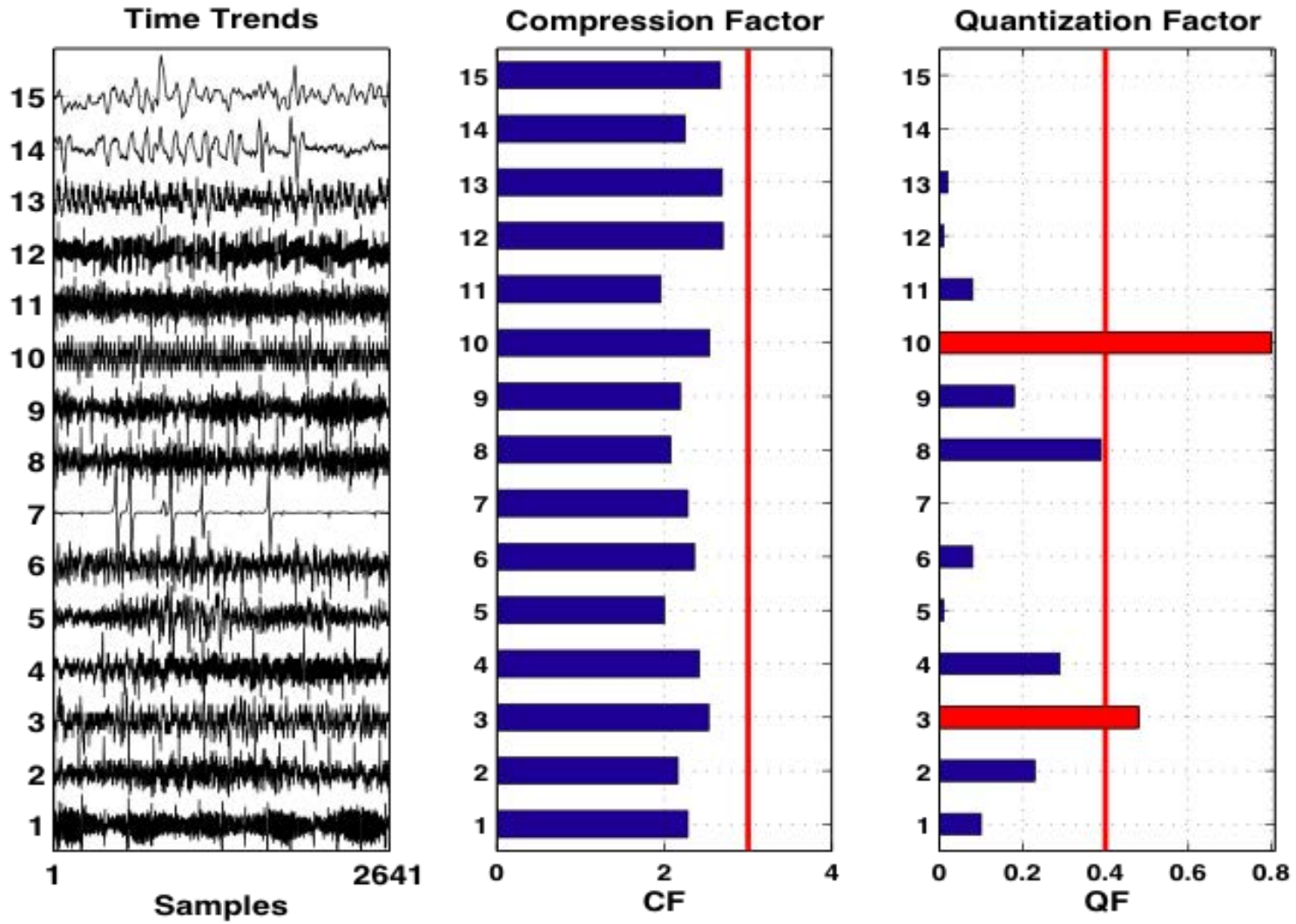
Correlation Color Plot of Reconstructed Data from Wavelet compression algorithm



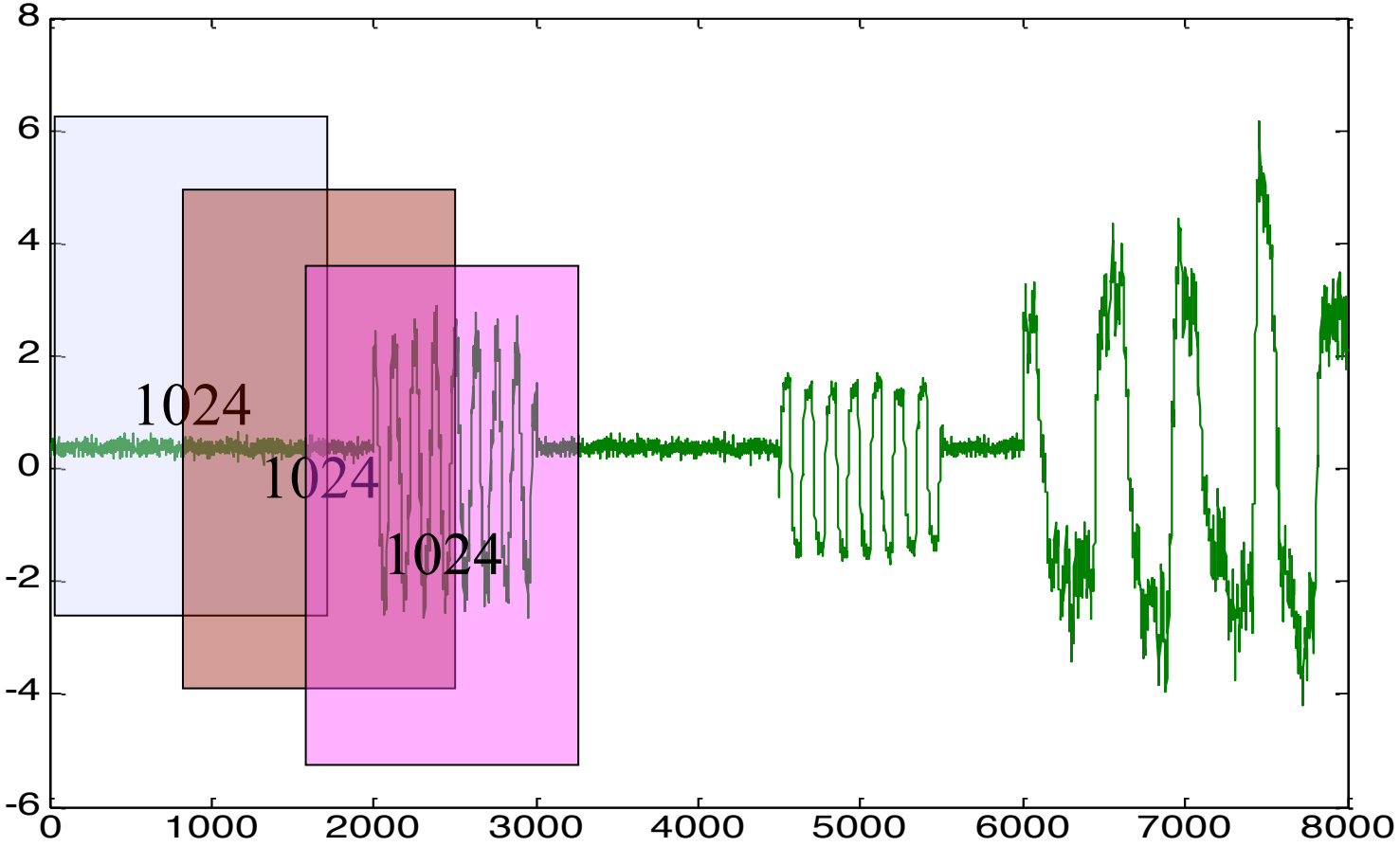
Correlation Color Plot of Reconstructed Data using PCAIA



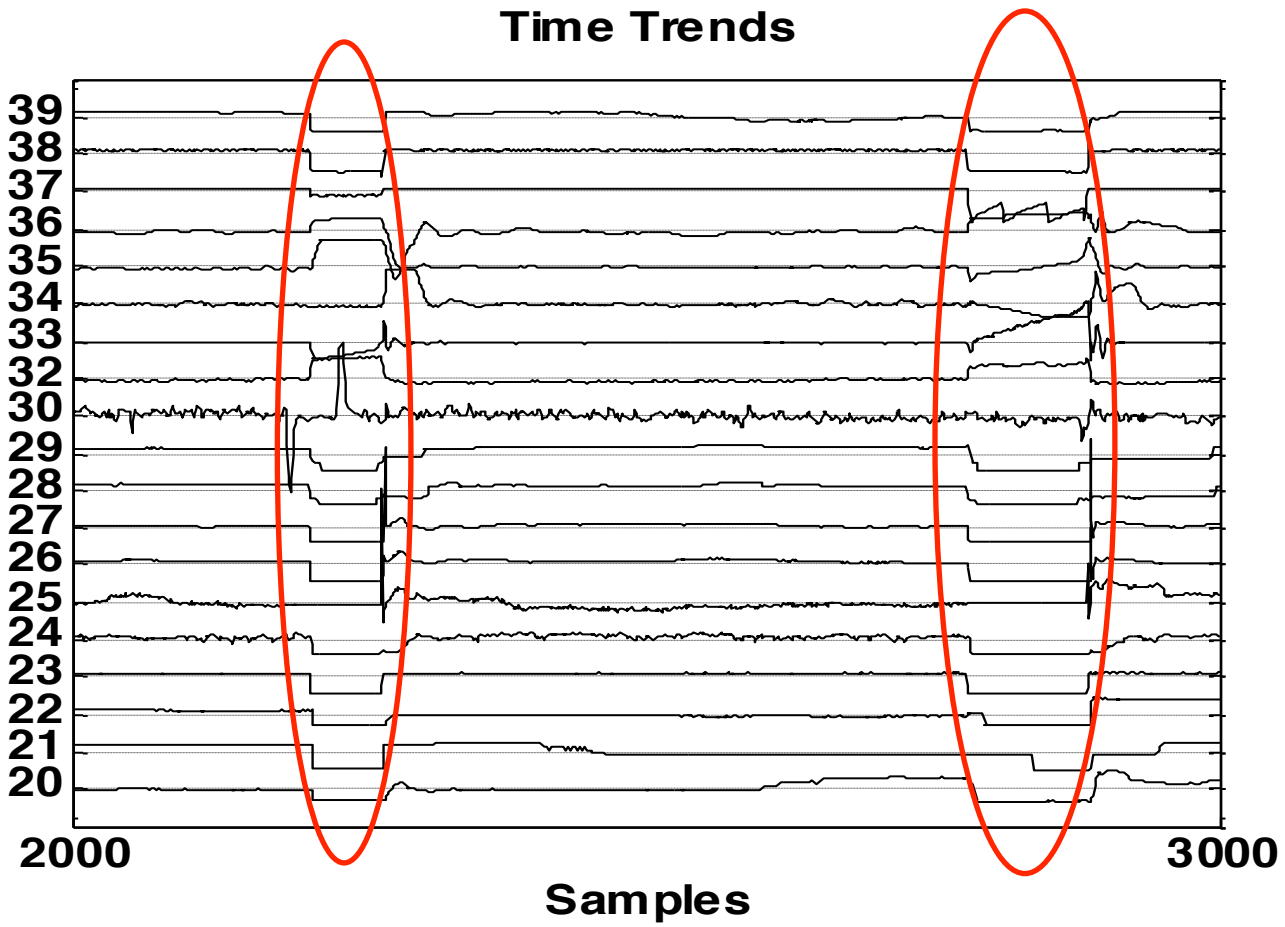
Compression & Quantization Analysis



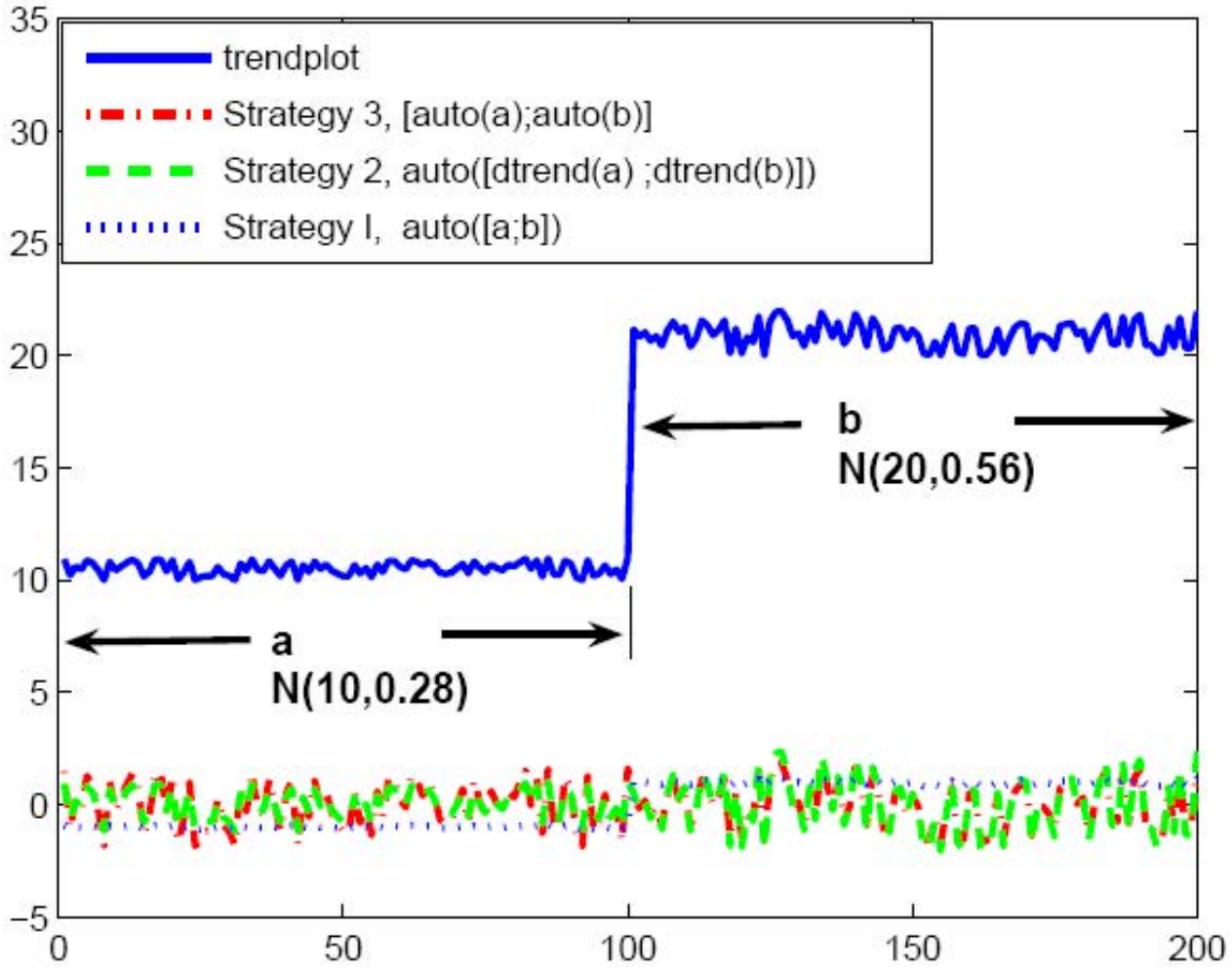
Challenges: Which Data Segment to use?



Data Quality: Segmentation



Data stitches



Concluding remarks

- Find useful patterns and information in data
- Use information to improve process performance:
  - Build models;
  - Monitor performance via KPIs;
  - Carry out FDD and root cause detection and analysis;
  - Leverage and maximize value in your data sitting in different silos;
  - Extract actionable information from your data.
- Data analytics should be a core engineering subject.