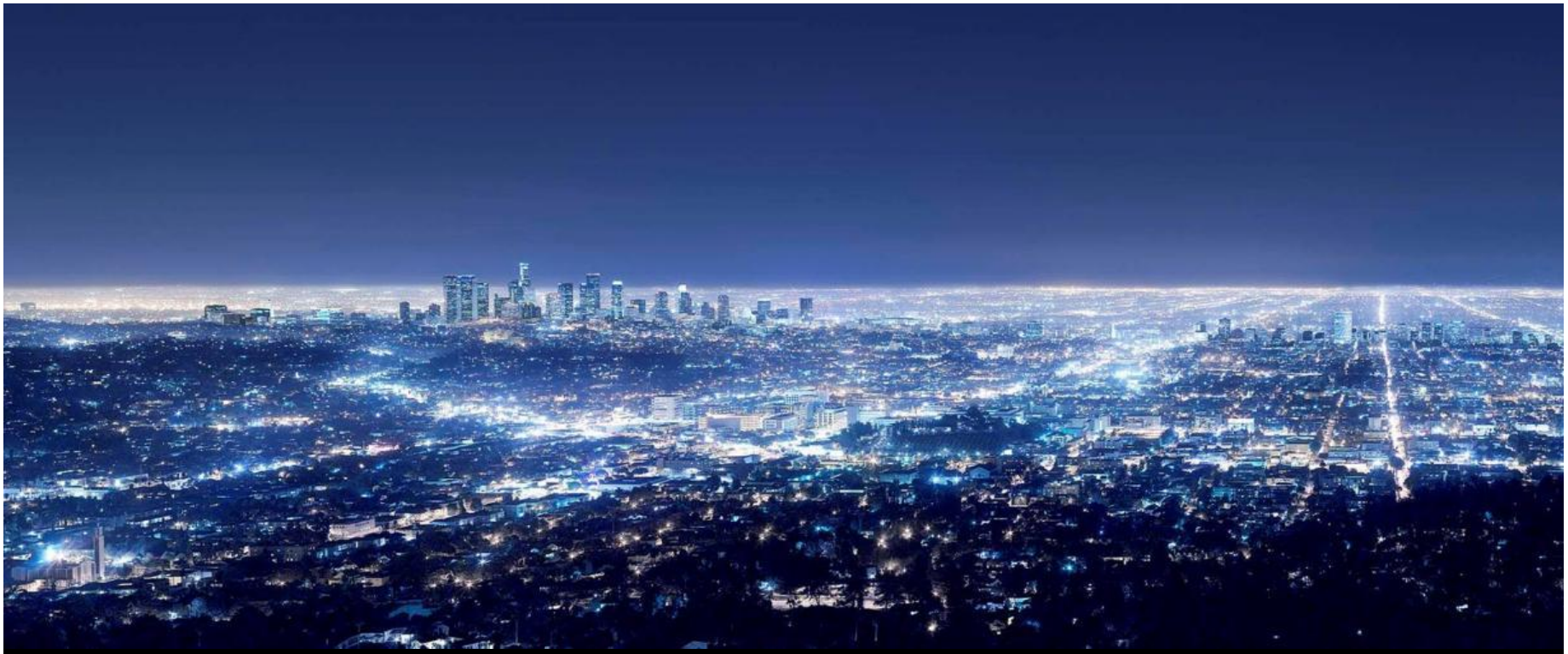


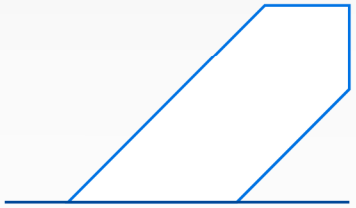
Ralf Gitzel, ABB – CBI 2016, 24.8.2016

Data Quality in Time Series Data
An Experience Report

Project Context and Overview

A global leader in power and automation technologies
Leading market positions in main businesses

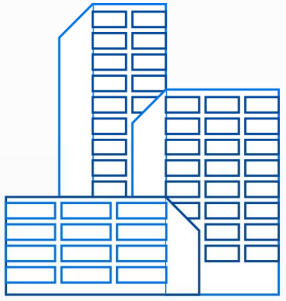
~150,000 employees



$42 billion
In revenue (2013)

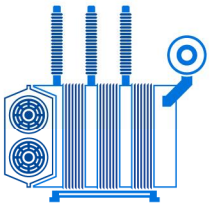
Present in +100 countries

Formed in 1988

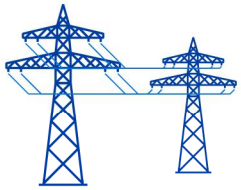


merger of Swiss (BBC, 1891) and Swedish (ASEA, 1883) engineering companies

How ABB is organized
Five global divisions



Power Products



Power Systems

Discrete Automation and Motion

Low Voltage Products

Process Automation

Project Context
Data Quality Analysis

- Goal is to establish a data quality library/process for all analytics projects within ABB
- Tool is being developed consecutively using data from different analytics projects

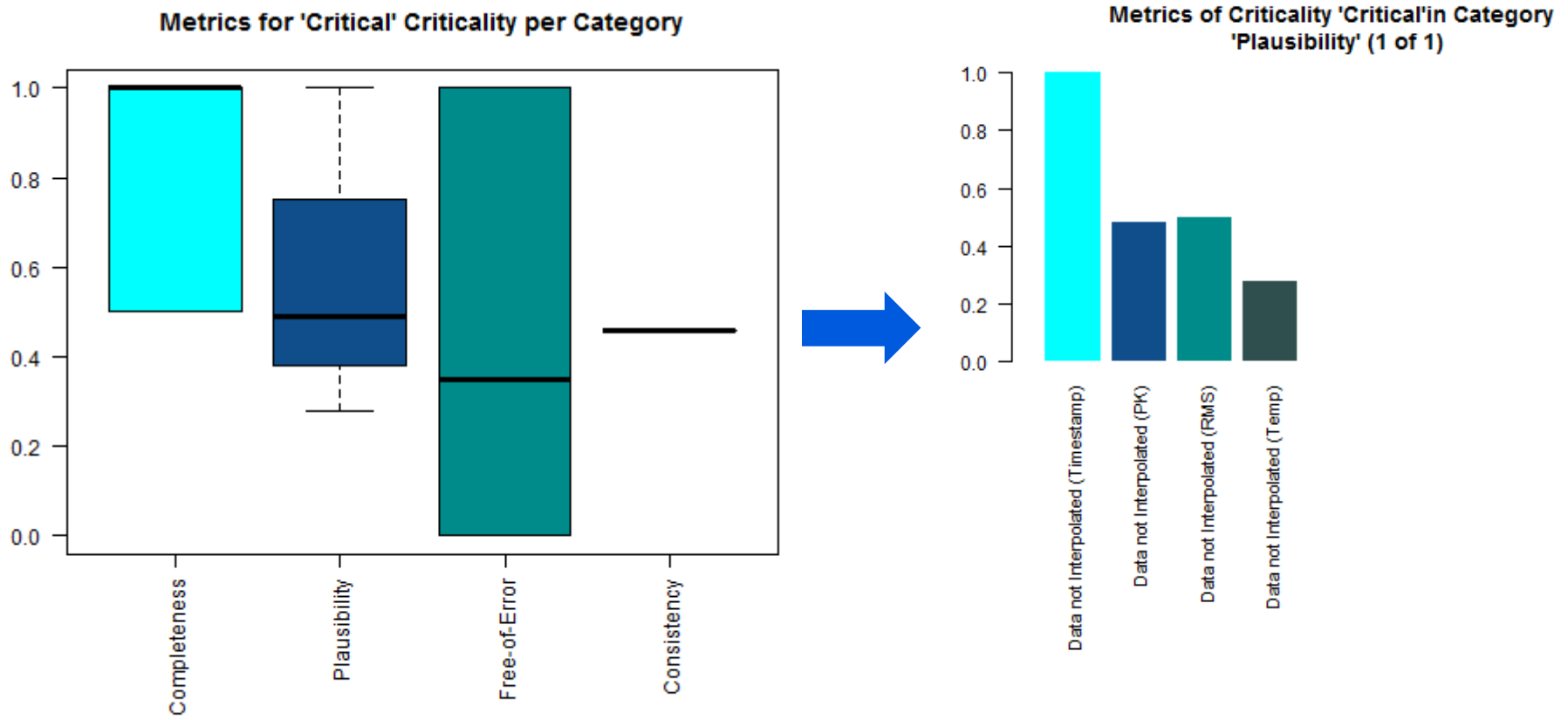
Key Time Series Data Quality Problems

Key Problems in Time Series
Literature Review

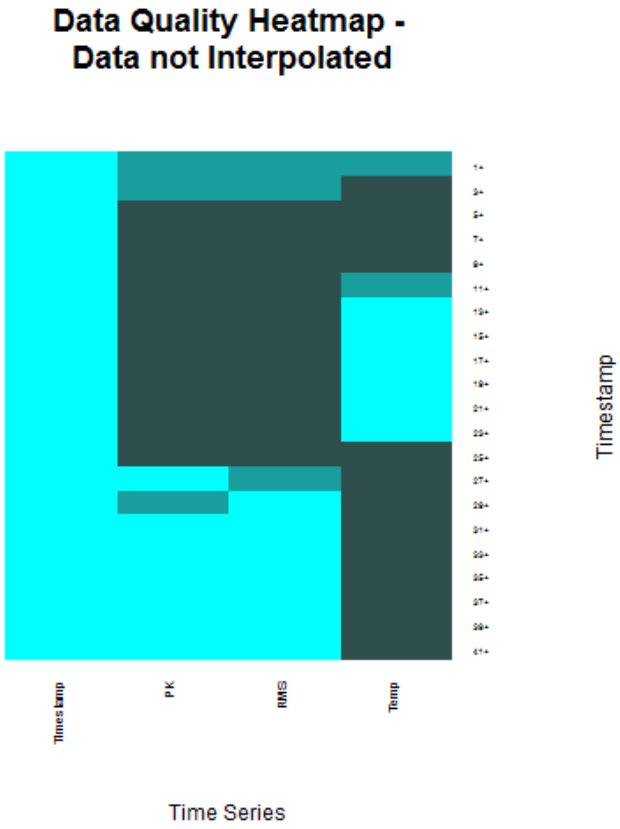
Problem Category	Example Problems
Missing Data	Empty fields
Obviously Wrong Data	Out of range, impossibly quick changes
Time Stamp Order	Unsynchronized clocks, delayed signals
Inaccurate Values	Measurement tolerance, rounded values, noise
Stuck Values	Data not updated
Implausible Values	Data that should be correlated is not, interpolated data is „just too good to be true“
….	*(and many more as described in the paper)*

Our Data Quality Prototype

Provide Quick Overviews
Top Level View with Drill-Down



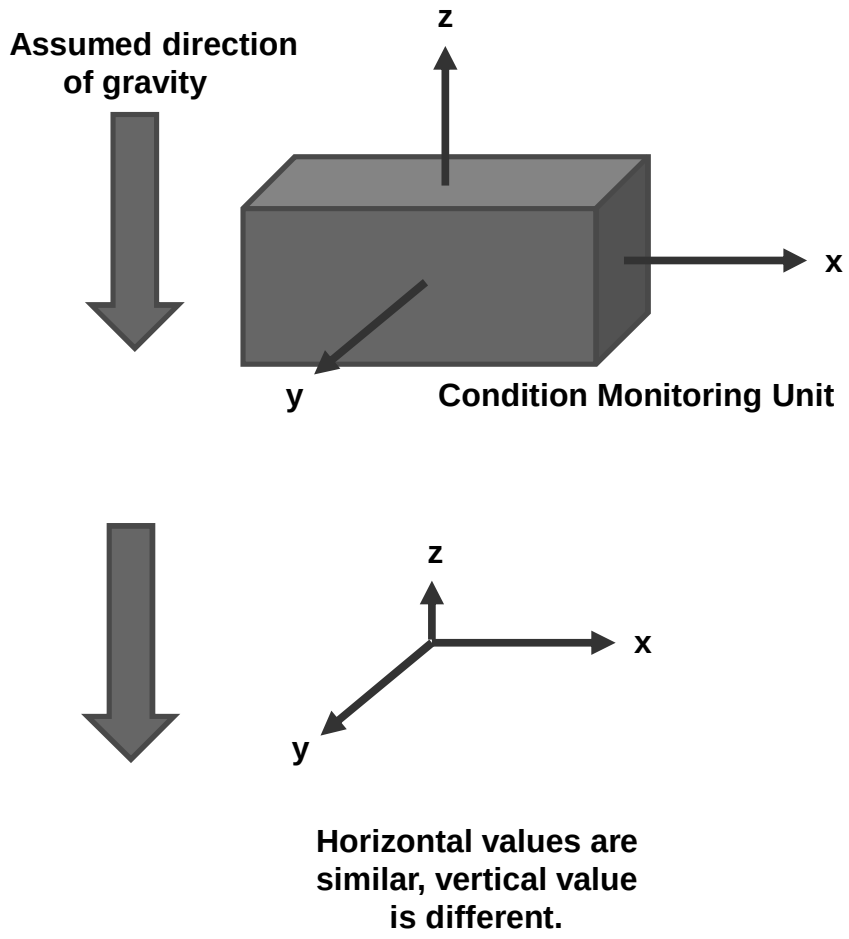
Provide Details
To Better Understand the Problems



Lessons Learned

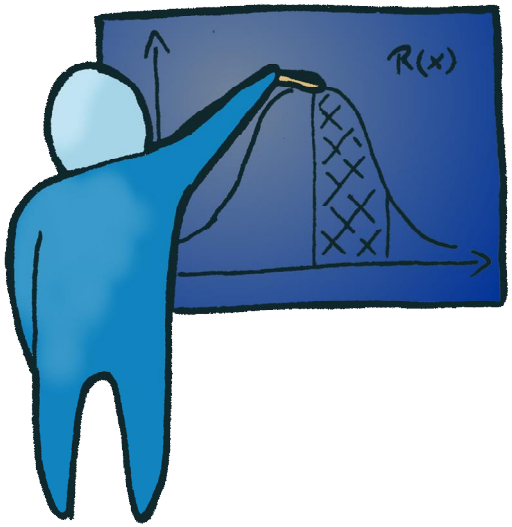
Lesson 1
Domain Knowledge is Important

- There is currently a trend in analytics to „let the data speak for itself“.
- However, without domain knowledge, valuable opportunities are wasted to identify problems in the data
- Example: Vibration of asset follows certain pattern if monitoring unit is applied in accordance with our assumptions.



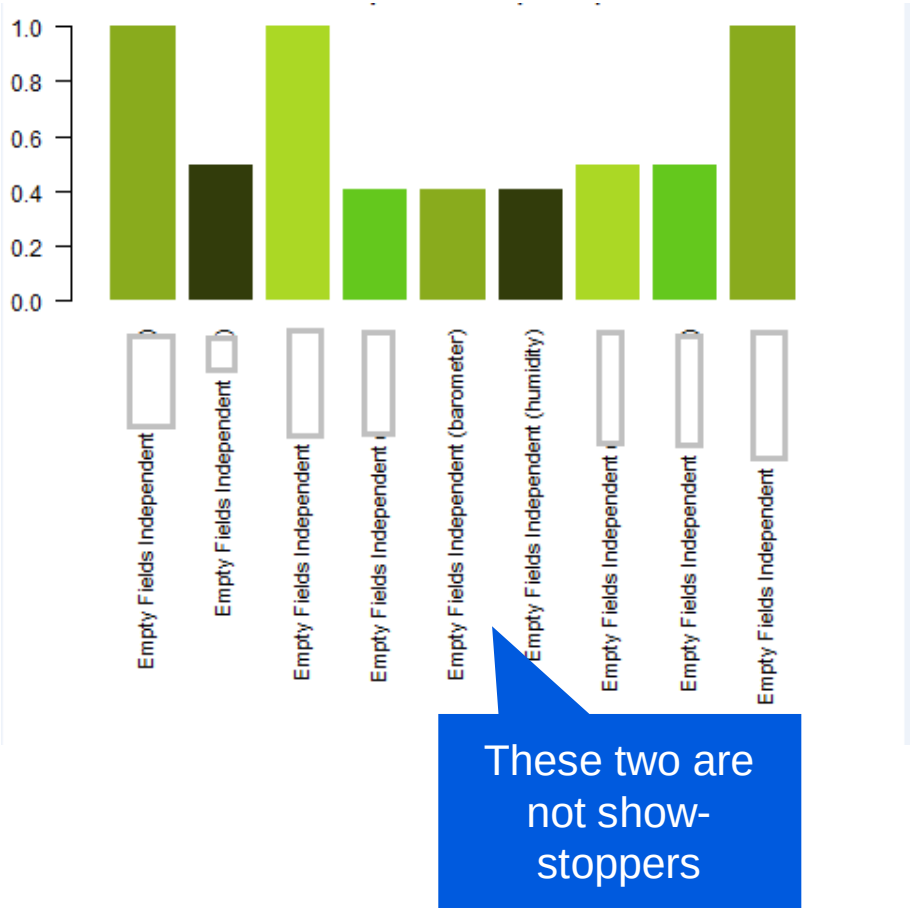
Lesson 2
Understand the Analytics

- Whether a data quality problem is important or not and how to best fix it, depends on the analytics
- Good example is missing data
  - If we aim to get a trend (increasing vibration), scattered missing values are not a problem
  - If we are looking for special events (e.g. spike in accelation), missing values might be highly problematic as they can hide these events. (Worse, there might be a connection between the spikes and the probability that the value is missing.)



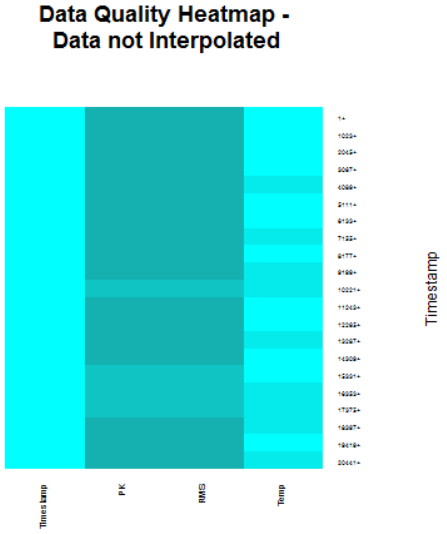
Lesson 3
Put the Most Pressing Issues on Top of the Stack

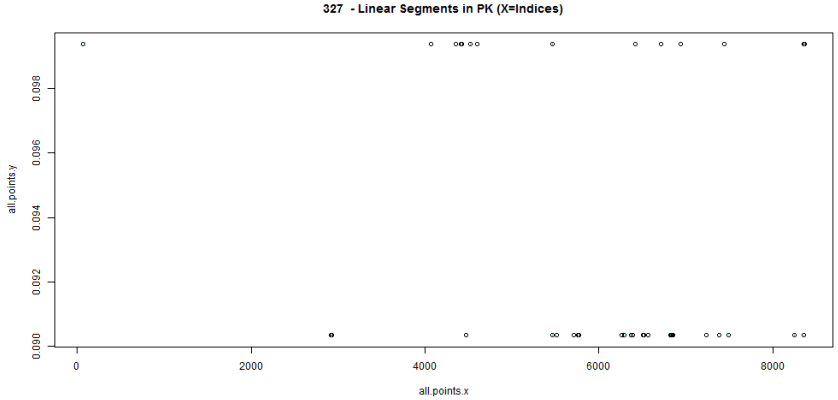
- It is easy to define metrics and to find (potential) problems
- However, not all problems have the same impact on the results of the analysis
  - Inaccurate data might lead to inaccurate results
  - Wrong data might lead to wrong conclusions
- It is important to understand the relevancy of problems and avoid hiding key issues under a haystack of minor issues



Lesson 4
Information Must be Actionable

- Metrics are fine but at the end of the day, we need to define actions
  - Which values are wrong?
  - Which areas are not useable?
- Example: 35% of all values in column x interpolated
  - Are there long stretches?
  - Are they concentrated in some spot?

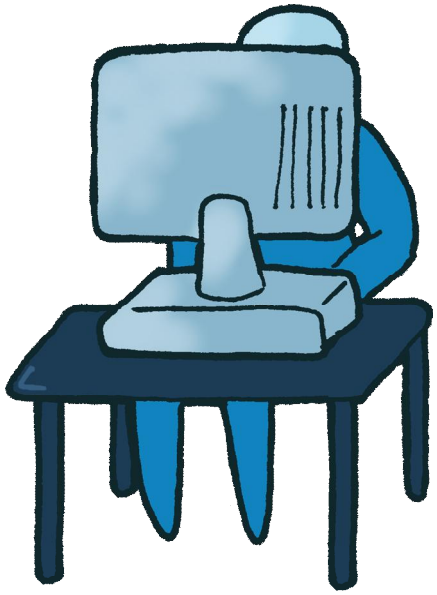




Lesson 5
The Devil is in the Details

- If we use packages not written by us, the way certain problems are handled can be confusing and time-consuming
- Example: Code snippet (right) did not parse a seemingly normal date
  - Explanation: We had the wrong timezone and due to daylight savings time 27.03.2016 2:00 never occured (clock skipped to 3:00, so there is a missing hour)
- Such puzzles take up time because the code crashes for no apparent reason

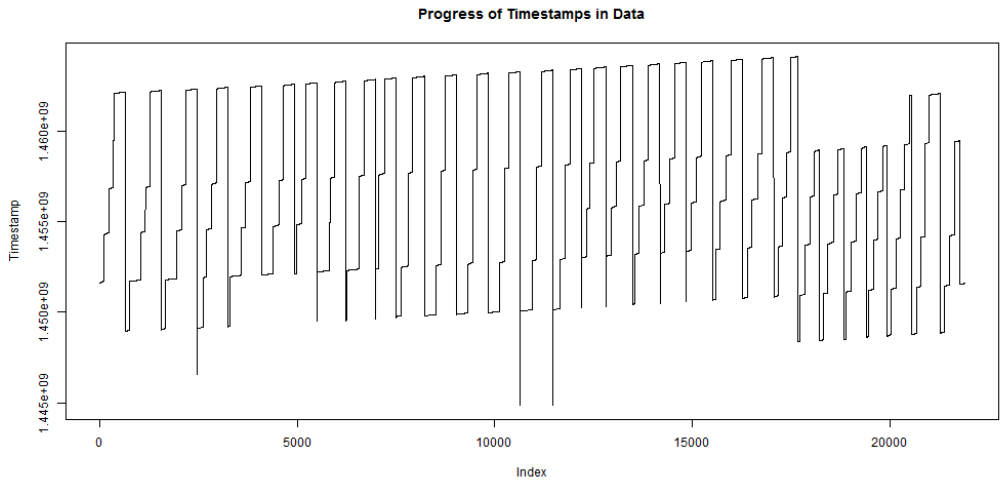
```
temp.date <- "27.03.2016 02:00"
temp.d <- as.POSIXlt(temp.date, "UTC", "%d.%m.%Y %H:%M")
as.numeric(temp.d) # returns NA

temp.date <- "27.03.2016 00:00"
temp.d <- as.POSIXlt(temp.date, "UTC", "%d.%m.%Y %H:%M")
as.numeric(temp.d) # returns 1459033200
```



Lesson 6
Some Problems will Slip By the Metrics

- You can create a list of problems that you find and maybe fix them
- However, this does not mean that everything else is correct
- Also, sometimes your choice of metric will downplay a problem
- Also, your data preparation can very well introduce new problems

Metric detected few violations but timestamp order is completely wrong



Summary

Summary
Data Quality Analysis

- Data Quality assessment is a bit like detective work
- it is easy to measure problems, it is hard
  - to find all problems
  - to focus on the important problems
- Domain and analytics algorithms influence what kind of problems we should be looking for – there is no one-size-fits-all data quality assessment
- Metrics can help to detect problems but human intervention is needed to avoid replacing one problem with another

Power and productivity
for a better world™