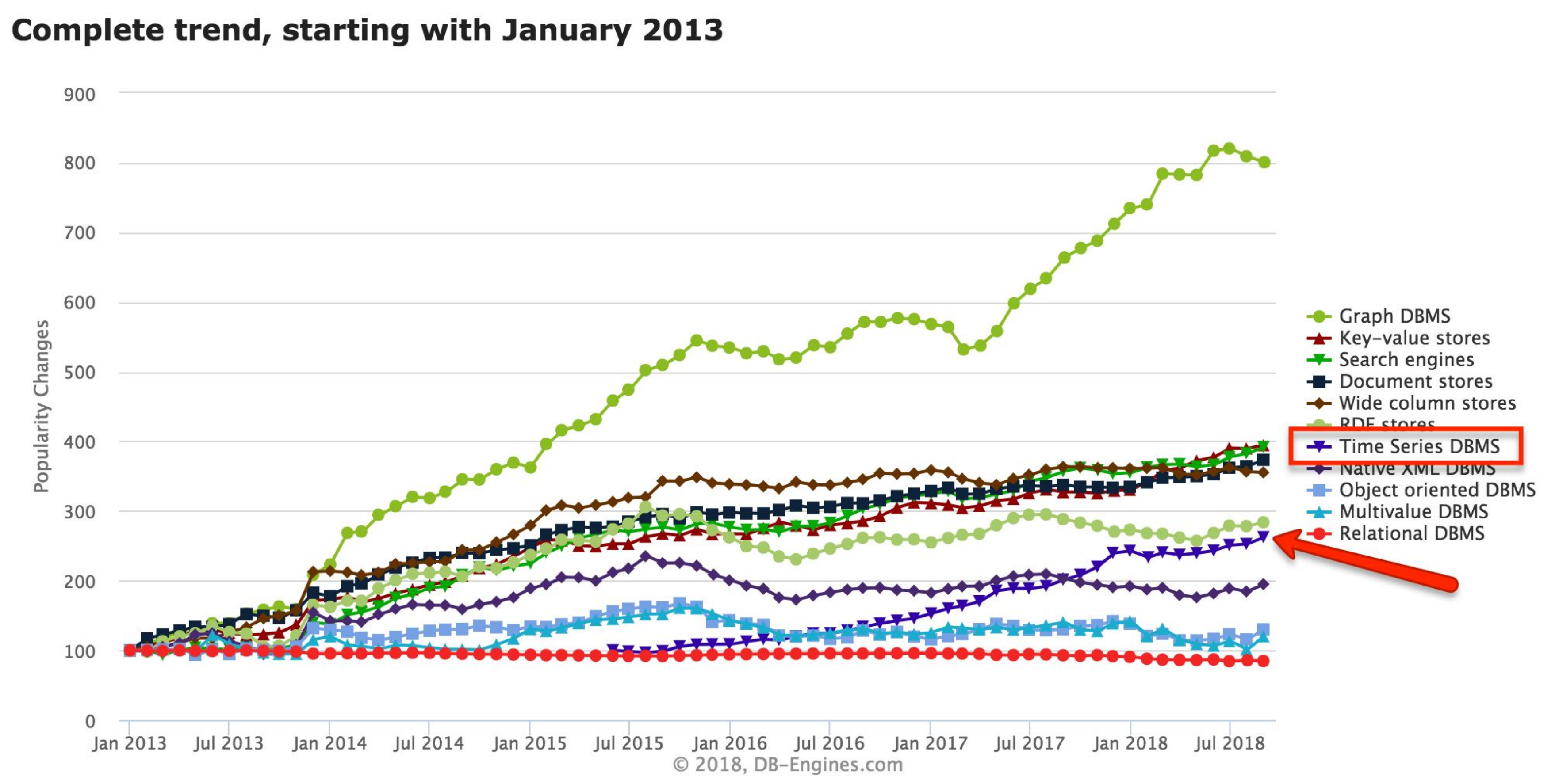
Time Series DB

Every data is time series ^_^

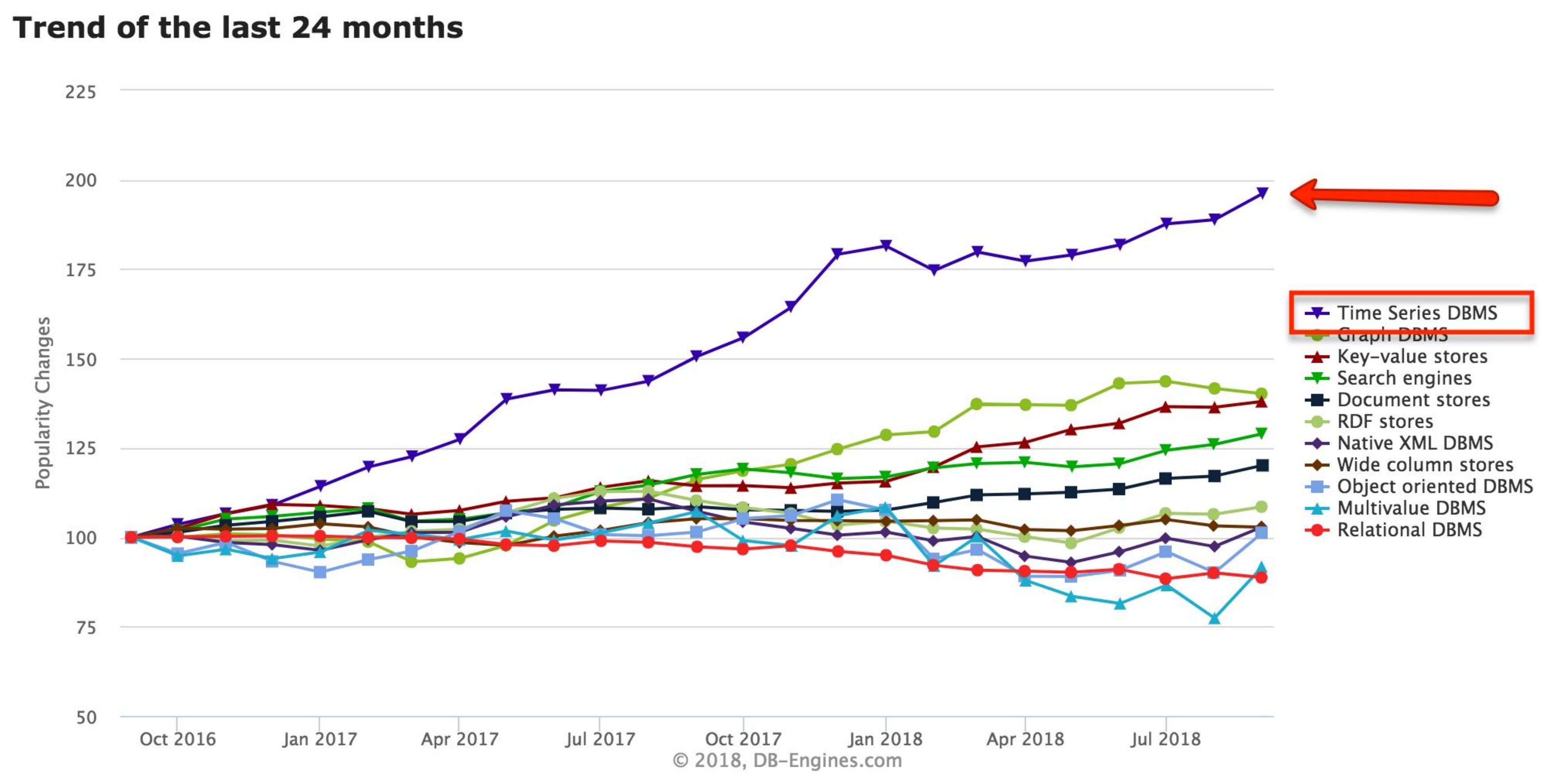
Is Time Series DB trandy?



Complete trend, starting with January 2013

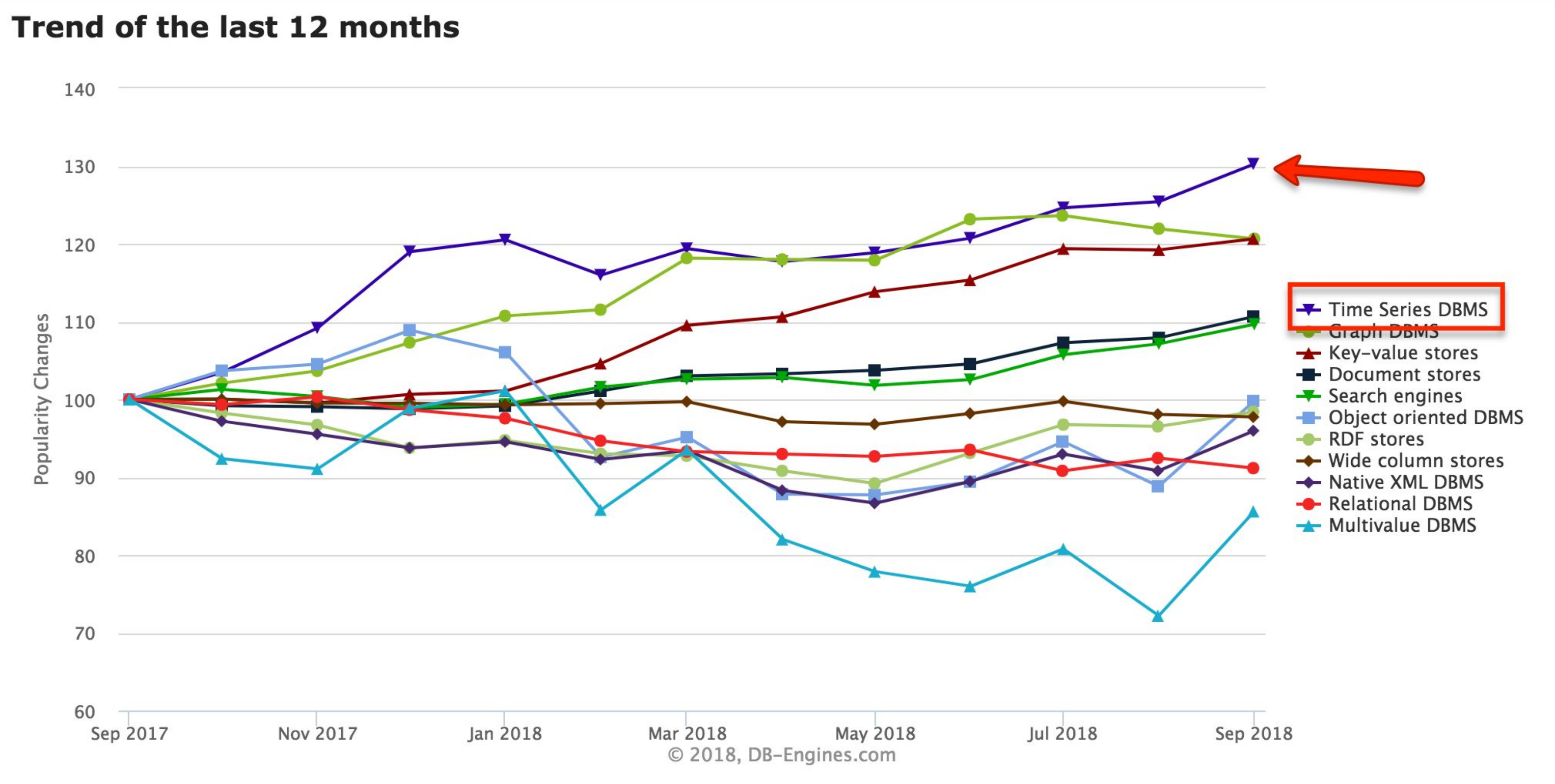
© 2018, DB-Engines.com

And 2 years!



Trend of the last 24 months
225
200
175
150
125
100
75
50
Popularity Changes
Oct 2016
Jan 2017
Apr 2017
Jul 2017
Oct 2017
Jan 2018
Apr 2018
Jul 2018
Time Series DBMS
Graph DBMS
Key-value stores
Search engines
Document stores
RDF stores
Native XML DBMS
Wide column stores
Object oriented DBMS
Multivalue DBMS
Relational DBMS
© 2018, DB-Engines.com

Lets check the last year!



How many time series DBs are there?

Influx = 15

Wikipedia = 6

Misframe = 50+

Really = ??

Why do we need Time Series DB?

To store Time Series Data!

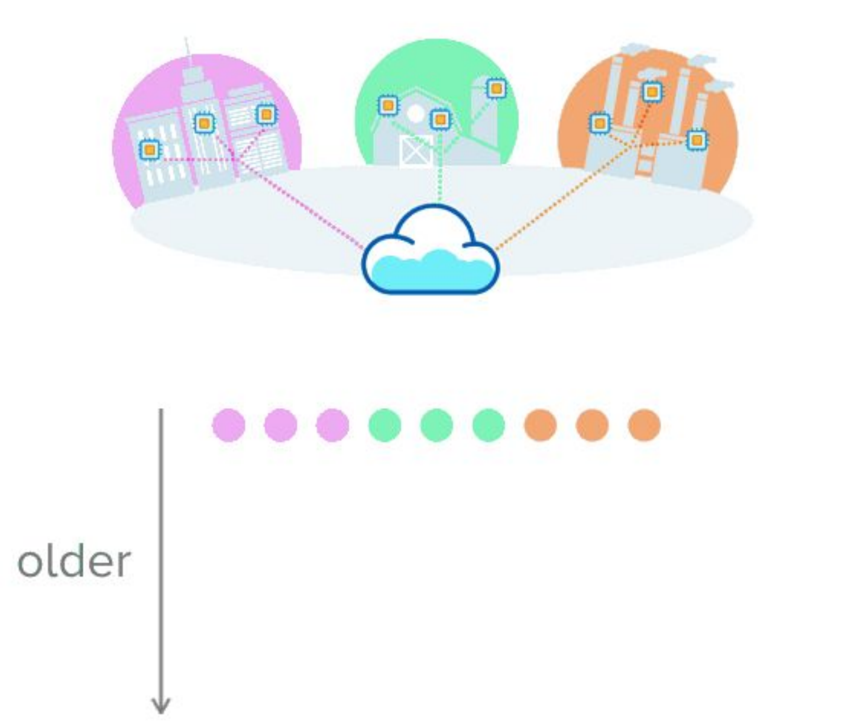
Thank you, captain Obvious.
Your work here is done!
ICANHASCHEEZBURGER.COM

Anyway its...

DATA SCIENCE
TIME
imgflip.com

What is Time Series Data?

An <u>ordered</u> sequence of values of a variable at equally spaced time intervals.

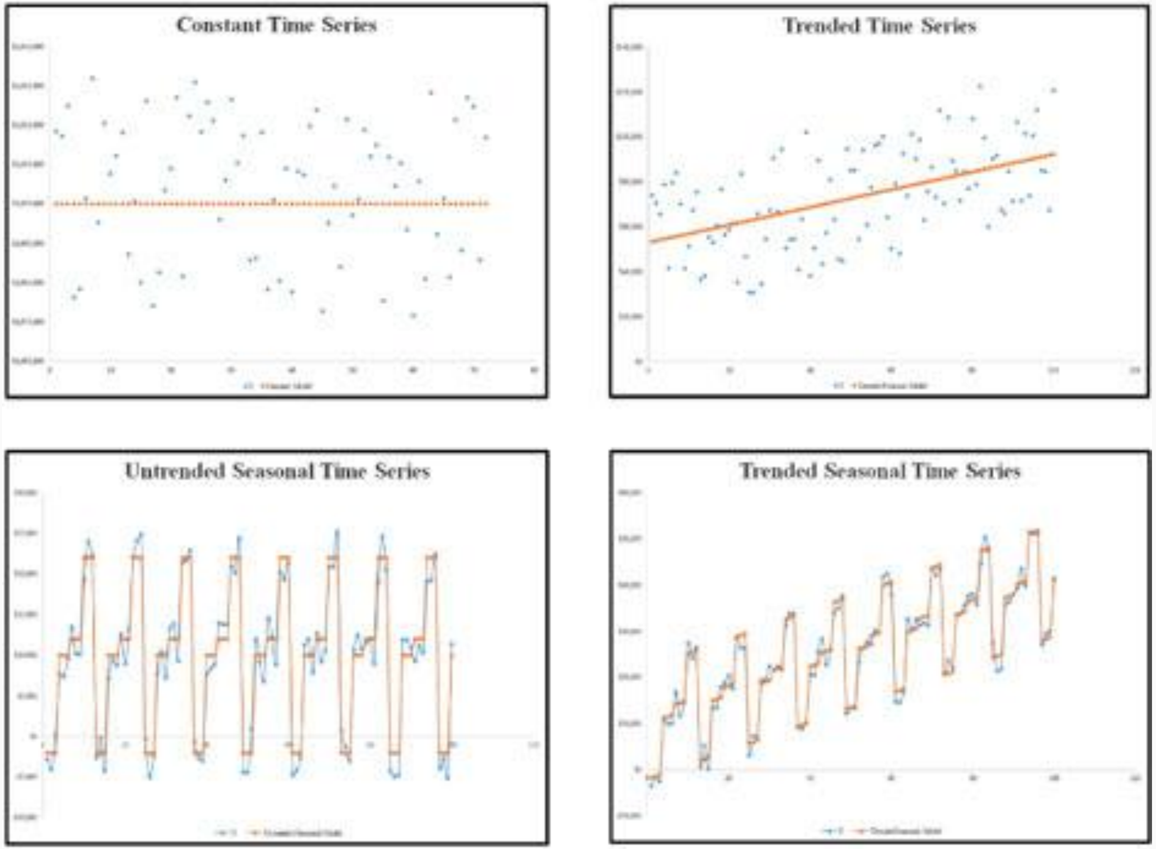


What is Time Series Data?

Time Series Data can be used for forcasting when values are related.

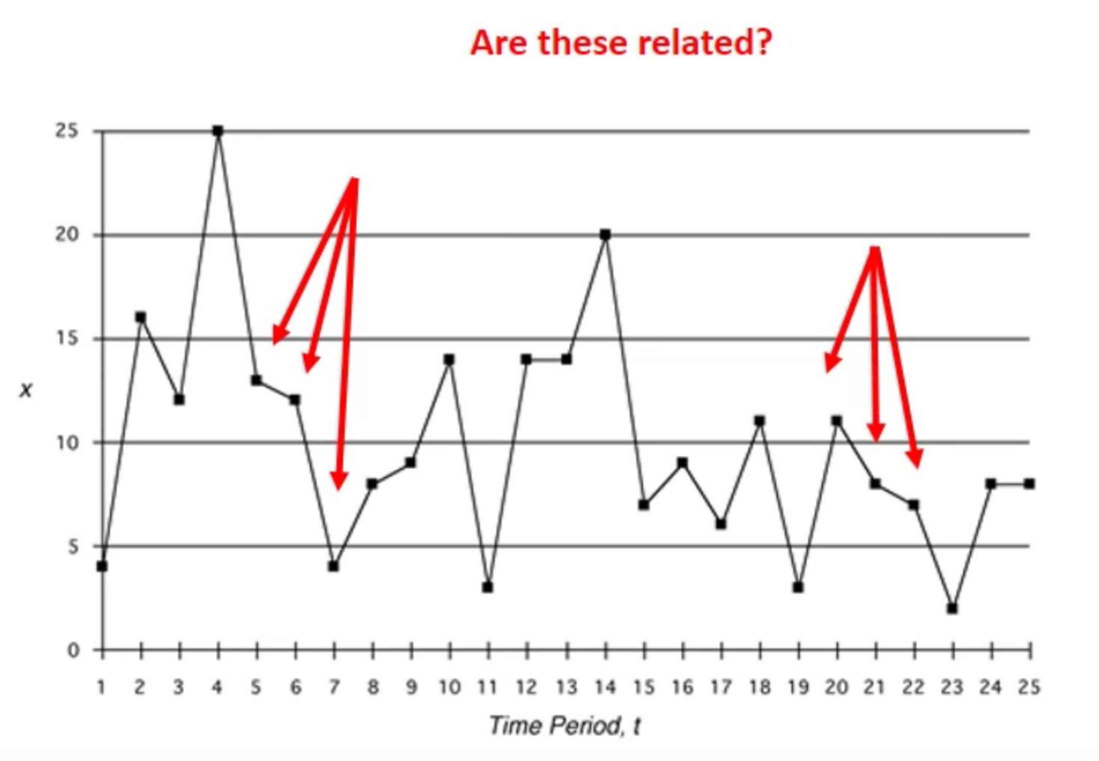
There are 4 types/patterns of related Time Series Data

Patterns Exhibited by Time Series

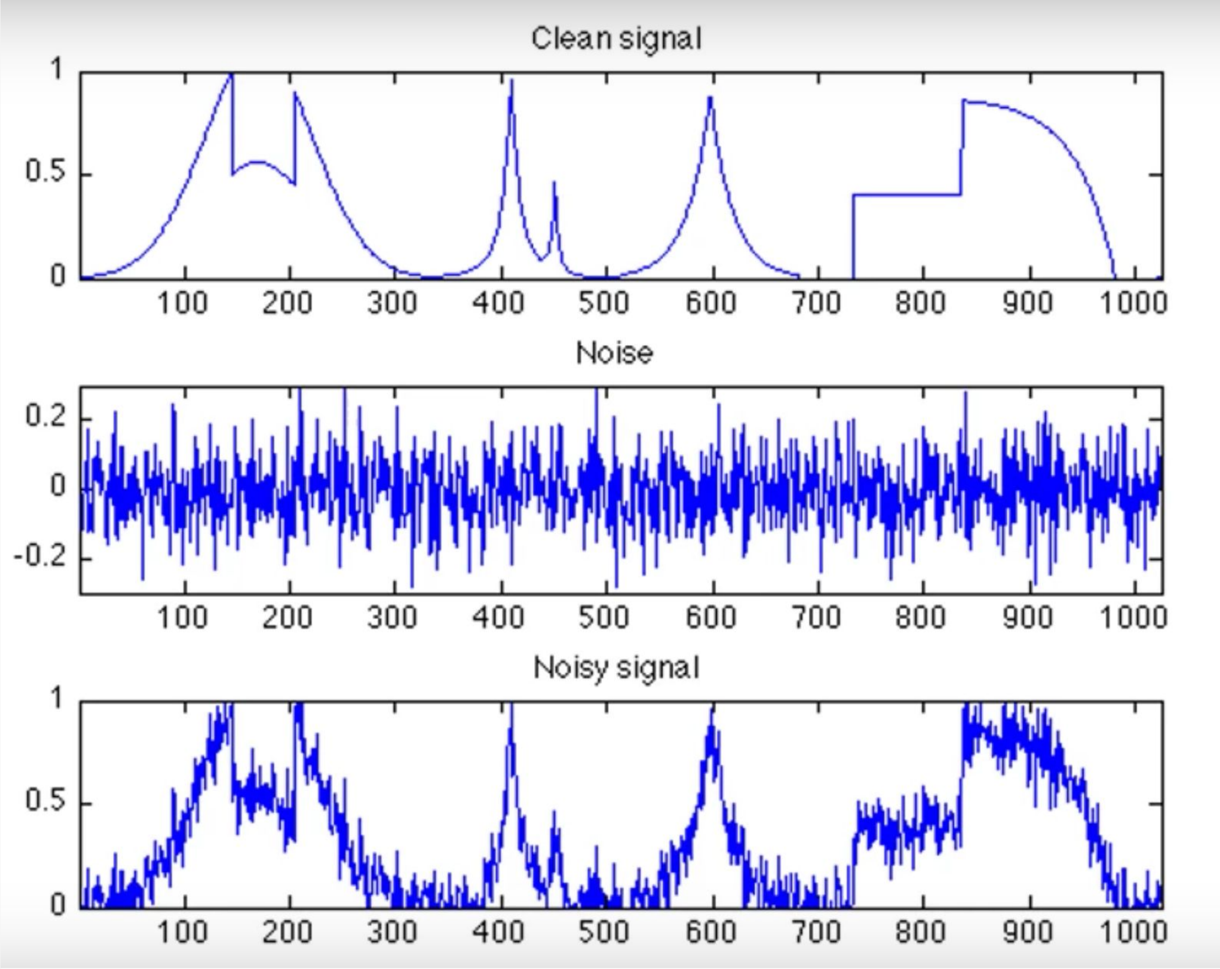


The data, which include random error, are in blue and the forecast models fitted to the data are in orange.

Is Time Series Data always related?



Is Time Series Data always related?

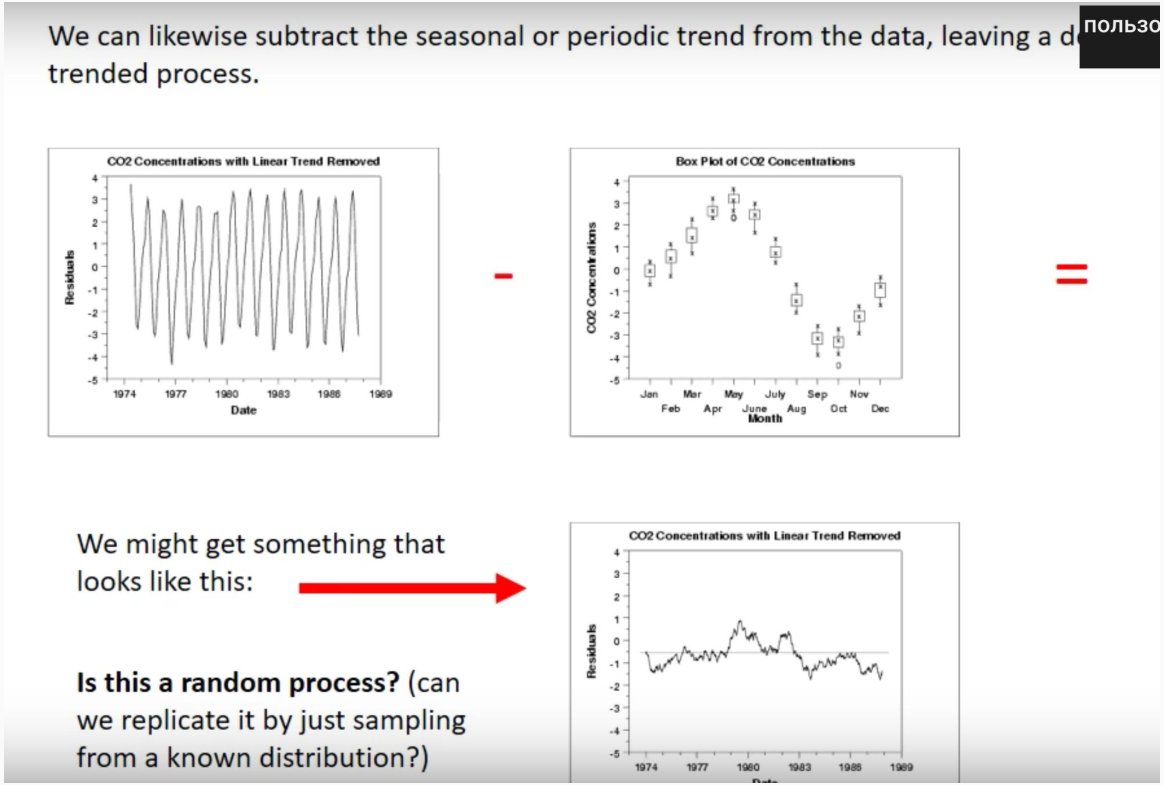


How do we deal with noise?

Noise can be filtered or smoothed.

How do we deal with noise?

Filtering

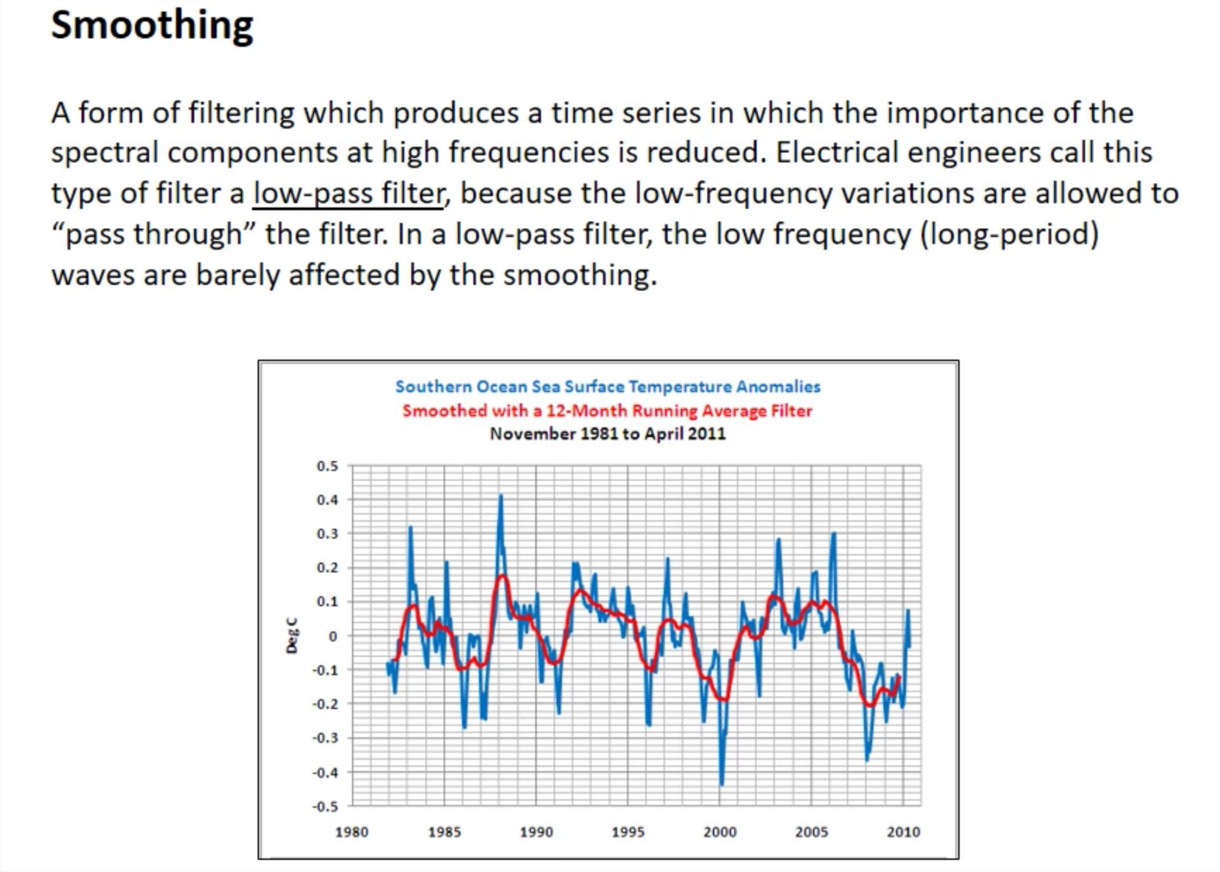


We can likewise subtract the seasonal or periodic trend from the data, leaving a d пользо trended process.

We might get something that looks like this:

Is this a random process? (can we replicate it by just sampling from a known distribution?)

How do we deal with noise?



Smoothing

A form of filtering which produces a time series in which the importance of the spectral components at high frequencies is reduced. Electrical engineers call this type of filter a low-pass filter, because the low-frequency variations are allowed to "pass through" the filter. In a low-pass filter, the low frequency (long-period) waves are barely affected by the smoothing.

Southern Ocean Sea Surface Temperature Anomalies
Smoothed with a 12-Month Running Average Filter
November 1981 to April 2011

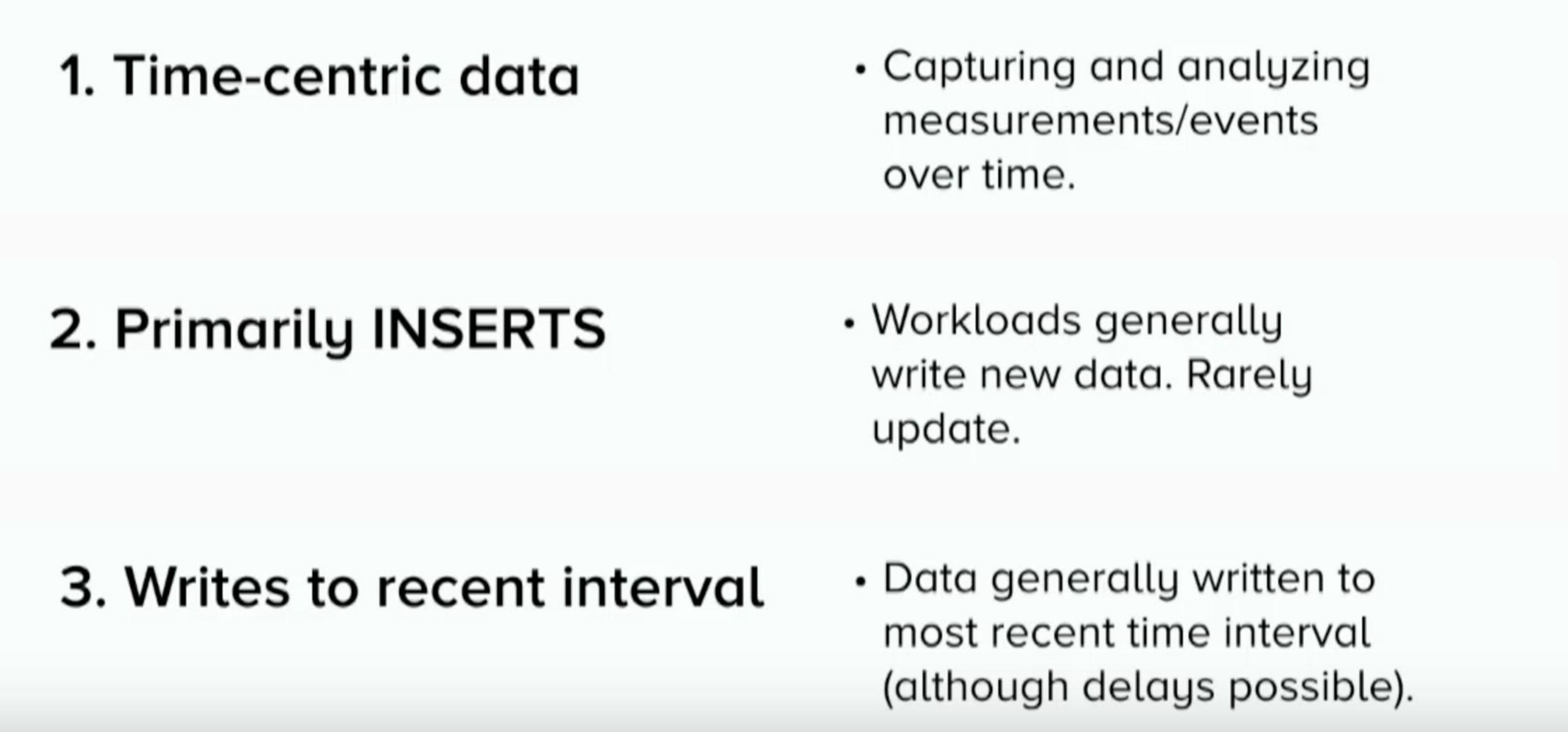
OK…

LESS DATA SCIENCE
TIME
imgflip.com

Time Series Data Examples

- Logs
- Metrics
- Financial Data
- Sensor Data
- Logistics Tracking
- Weather Data
- etc.

Time Series data has 3 characterisitcs

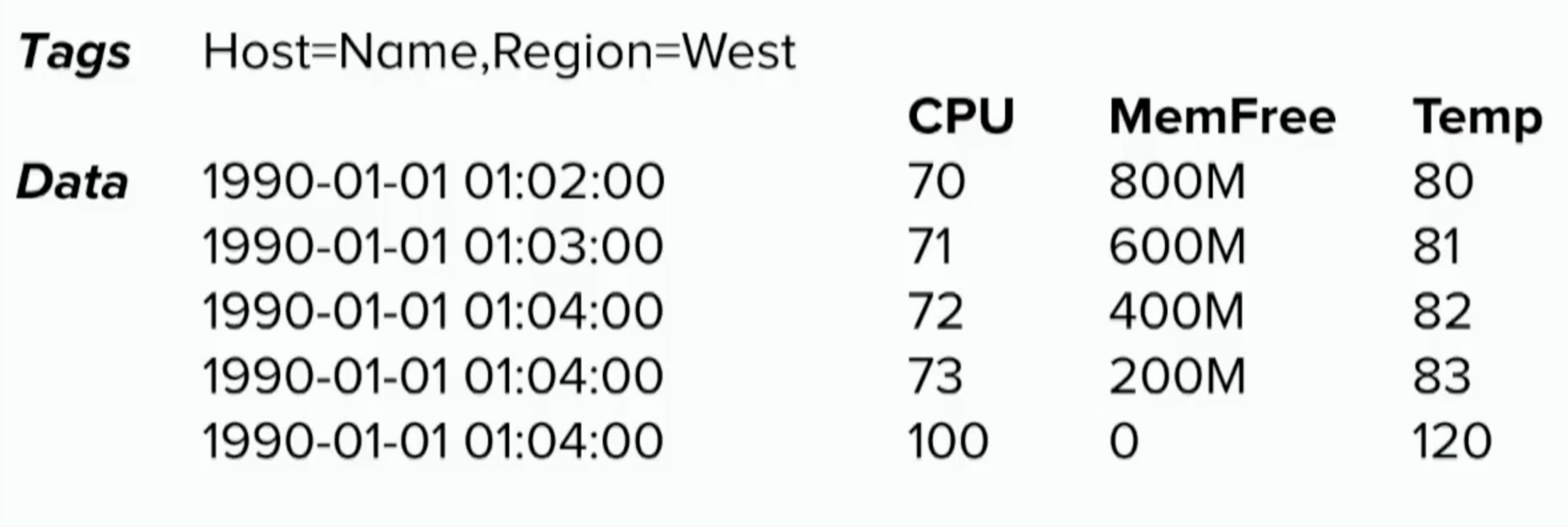


1. Time-centric data
   - Capturing and analyzing measurements/events over time.

2. Primarily INSERTS
   - Workloads generally write new data. Rarely update.

3. Writes to recent interval
   - Data generally written to most recent time interval (although delays possible).

Simple example of time series data



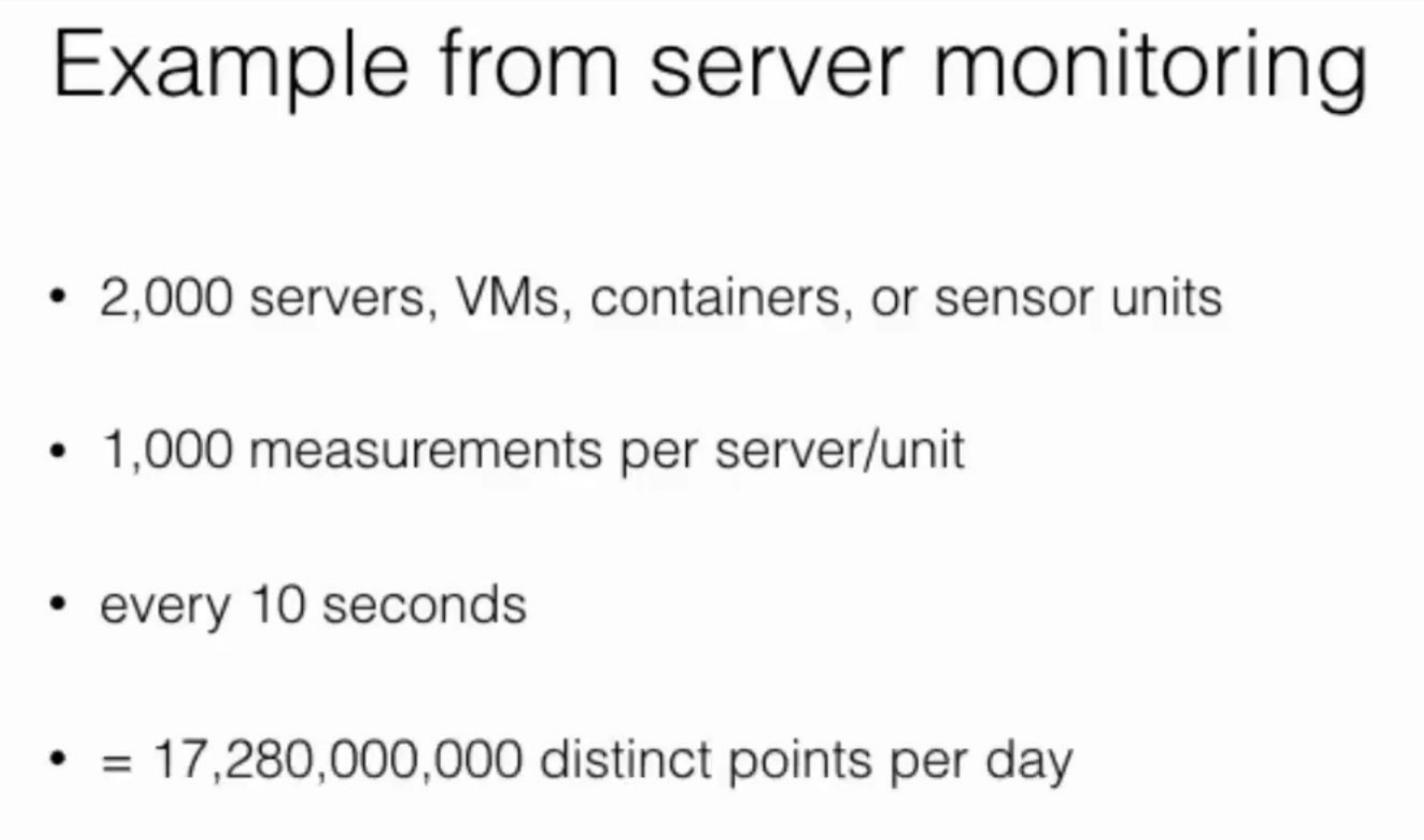
Tags	Host=Name,Region=West			
		CPU	**MemFree**	**Temp**
Data	1990-01-01 01:02:00	70	800M	80
	1990-01-01 01:03:00	71	600M	81
	1990-01-01 01:04:00	72	400M	82
	1990-01-01 01:04:00	73	200M	83
	1990-01-01 01:04:00	100	0	120

So can we use traditional RDBMS?

Yes

Till some point

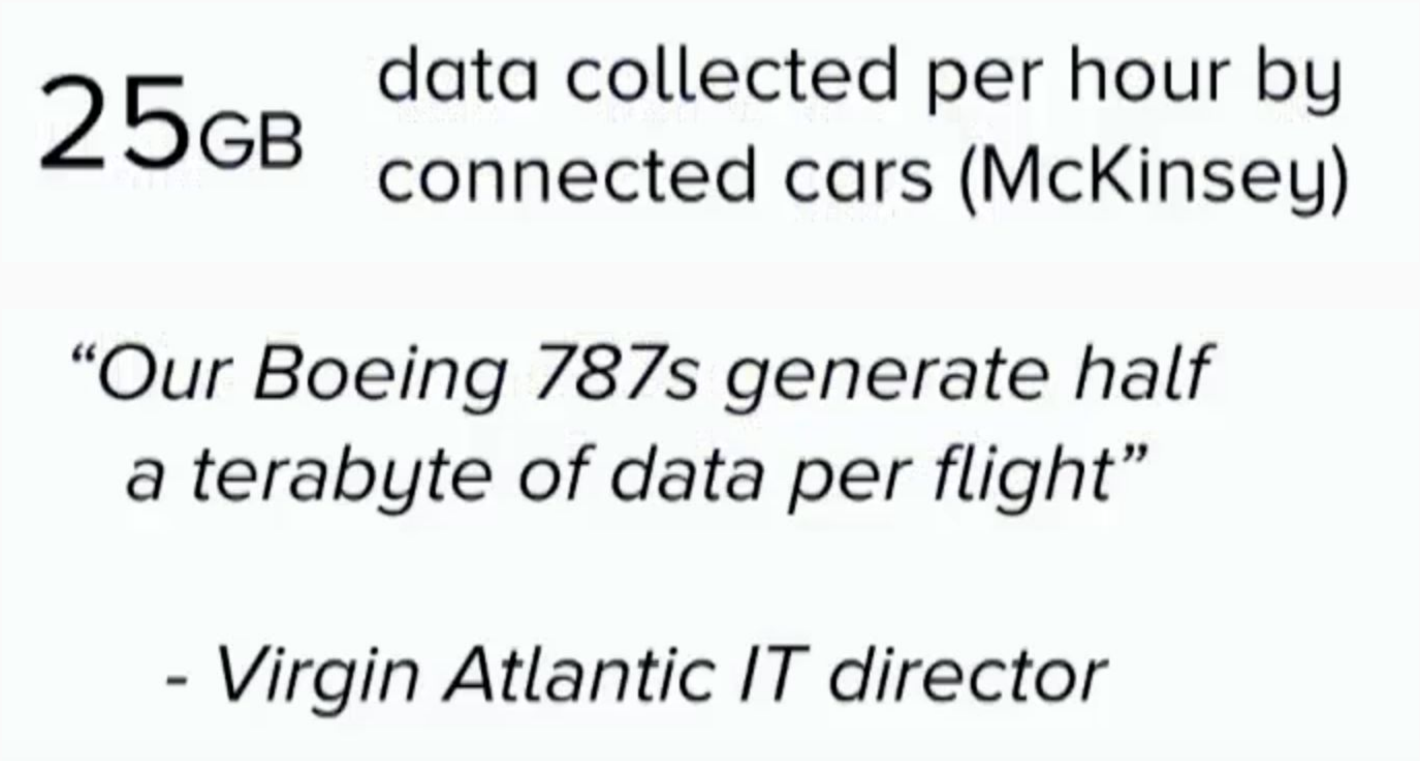
So can we use traditional RDBMS?



Example from server monitoring

- 2,000 servers, VMs, containers, or sensor units
- 1,000 measurements per server/unit
- every 10 seconds
- = 17,280,000,000 distinct points per day

So can we use traditional RDBMS?



25GB data collected per hour by connected cars (McKinsey)

“Our Boeing 787s generate half a terabyte of data per flight”

- Virgin Atlantic IT director

When Time Series DBs are needed?

Scaling

Compression

Downsampling (summaries)

Fast range queries

Timestamp as index

Aging out data

When Time Series DBs are **not** needed?

Not time series data.

Text agregation.

1 line reads. Range is better.

Weak Join functionality?!

Updates are expensive.

Other NoSQLs can do that. Why do we still need Time Series DBs?

Influx Data compared compared their TSDB with Cassandra, MongoDB and HBase

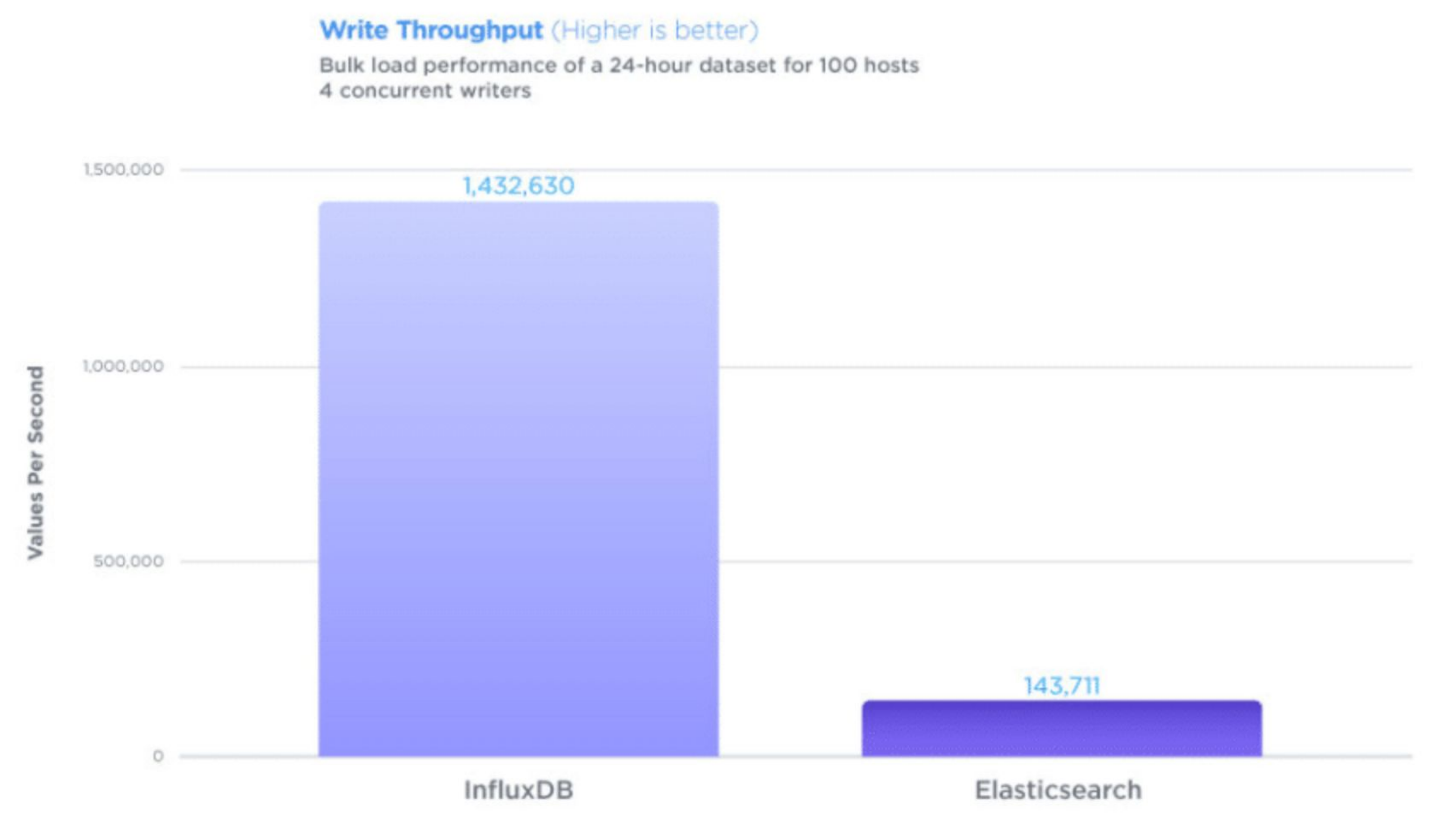
Influx Data vs others

TSDBs require less development effort.

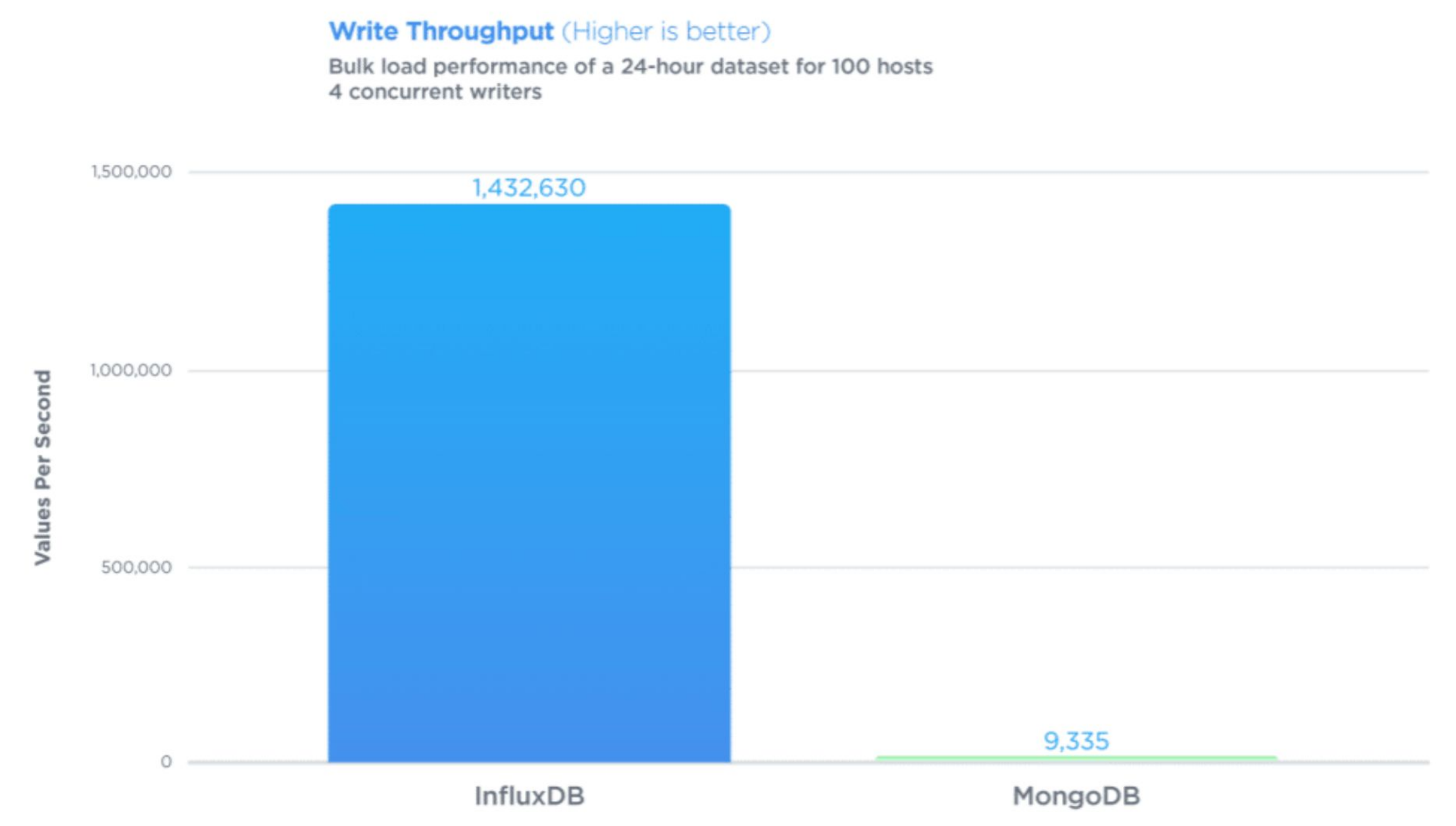
TSDBs do not require special CRUD API.

Influx Data TSDB does not require additional monitoring, alerting and visualization tools.

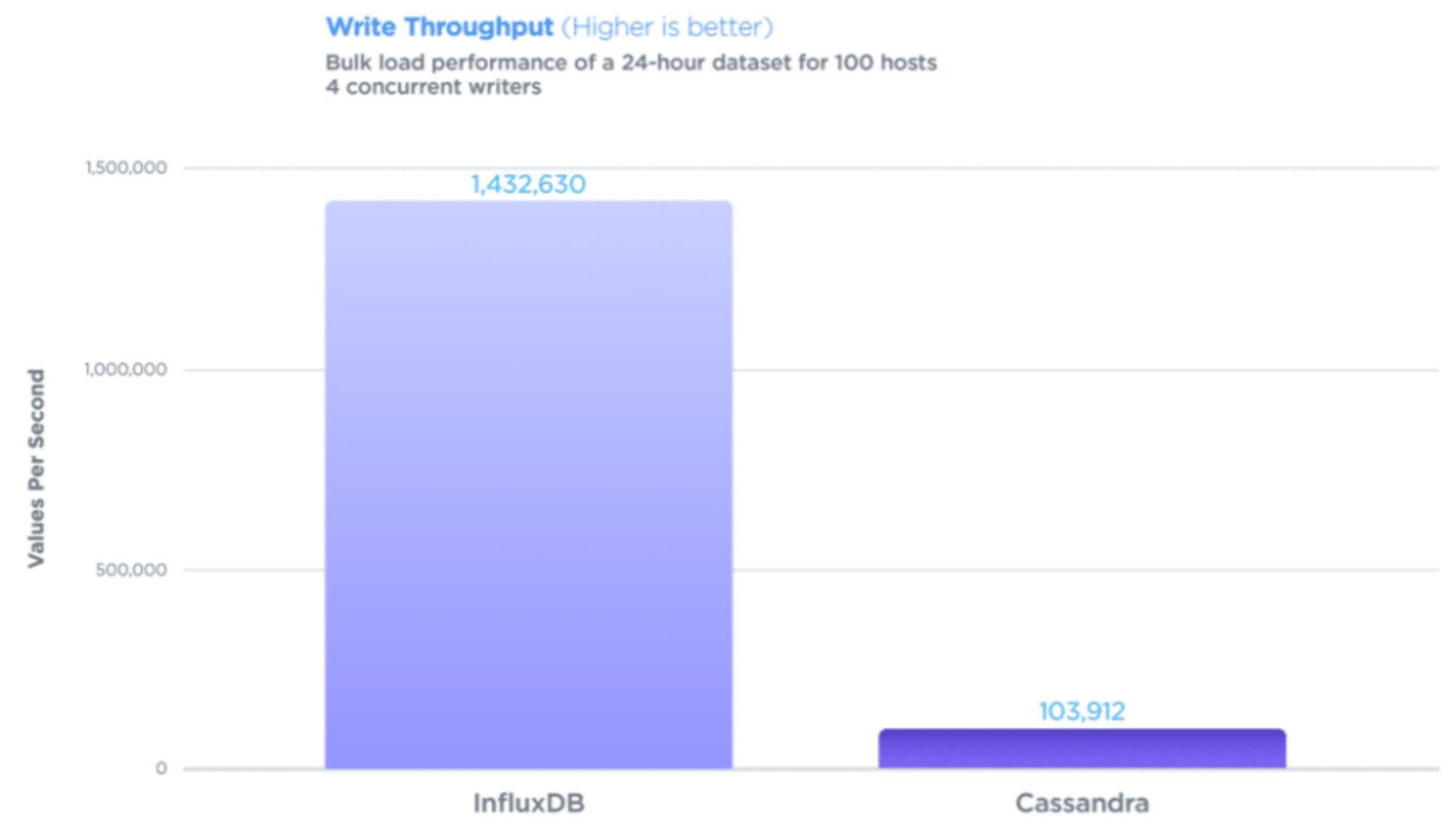
Influx Data TSDB vs ElasticSearch



Influx Data TSDB vs MongoDB



Influx Data TSDB vs Cassandra



Write Throughput (Higher is better)

Bulk load performance of a 24-hour dataset for 100 hosts
4 concurrent writers

Really?

I had to check simple scenario: InfluxDB vs. Elasticsearch

1. Find ready to use docker images
2. Use the simplest tutorials, without any hacks that improve performance
3. Write simple rows with random numbers and timestamps
4. Use jMeter to write data in 4 threads 50_000 inserts each
5. On the same machine (my laptop)
6. Reboot machine before every test
7. Measure results

InfluxDB API

```
# Create DB
curl -XPOST http://localhost:8086/query --data-urlencode "q=CREATE DATABASE mydb"

# Insert data
curl -i -XPOST 'http://localhost:8086/write?db=mydb&precision=ms' \
                    --data 'kinda,tag=test,thread=1 randomValue=42 1537701253843'
```

Elasticsearch API

```
# Create DB
# Meh

# Insert data
curl -XPOST http://localhost:9200/mydb/kinda \
        -H 'Content-Type: application/json' \
        --data '{"tag":"test", "thread":1, "randomValue":42,  \
        "timestamp":1537701253843'}'
```

Results ...

Results
(Use 4 threads to insert 50_000 records each)

	Elasticsearch	**InfluxDB**
Avg write/s	548	1132
Avg write ms	7	3
Total duration m	06:04	02:56

References

1. What the heck is time-series data? - https://www.youtube.com/watch?v=7hxXU9dceaE
2. Time Series Analysis - https://www.youtube.com/watch?v=Prpu_U5tKkE
3. Time Series Database Lectures #1 !! - https://www.youtube.com/watch?v=2SUBRE6wGiA
4. List of Time Series Data Bases - https://misfra.me/2016/04/09/tsdb-list/
5. Time Series DBs vs. other DBs - https://www.influxdata.com/time-series-database/
6. Really? - https://github.com/alex-d-bondarev/learn-timescale

That's all Folks!