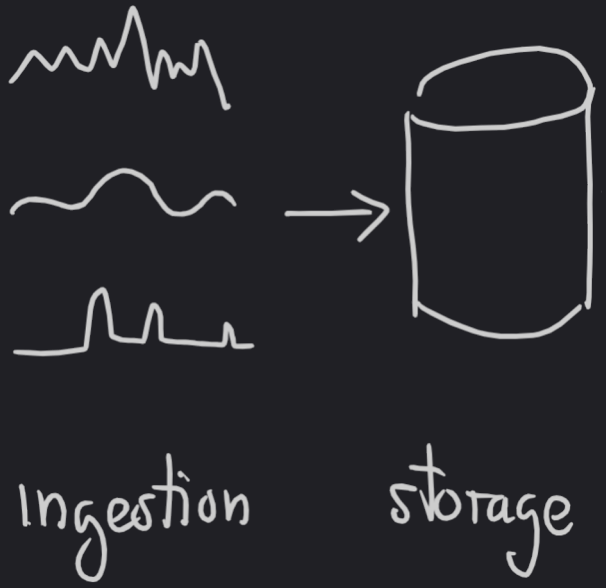
Data Ingestion

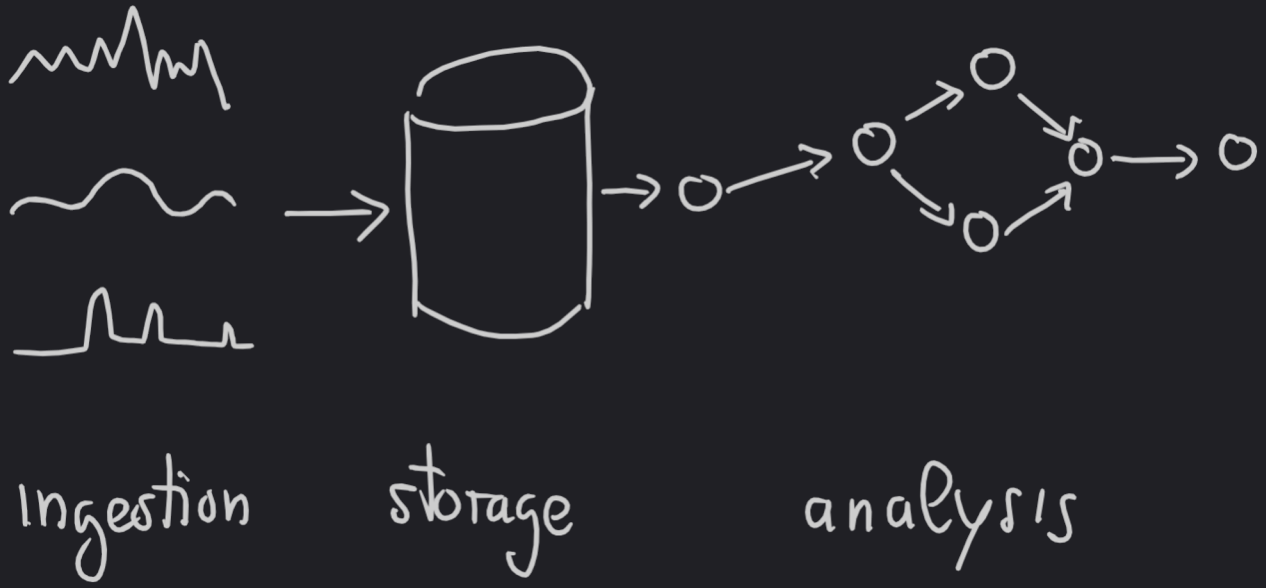
Emanuele Della Valle
Prof. @ Politecnico di Milano
Founder & Partner @ Quantia Consulting
Marco Balduini
Founder & CEO @ Quantia Consulting

Introduction

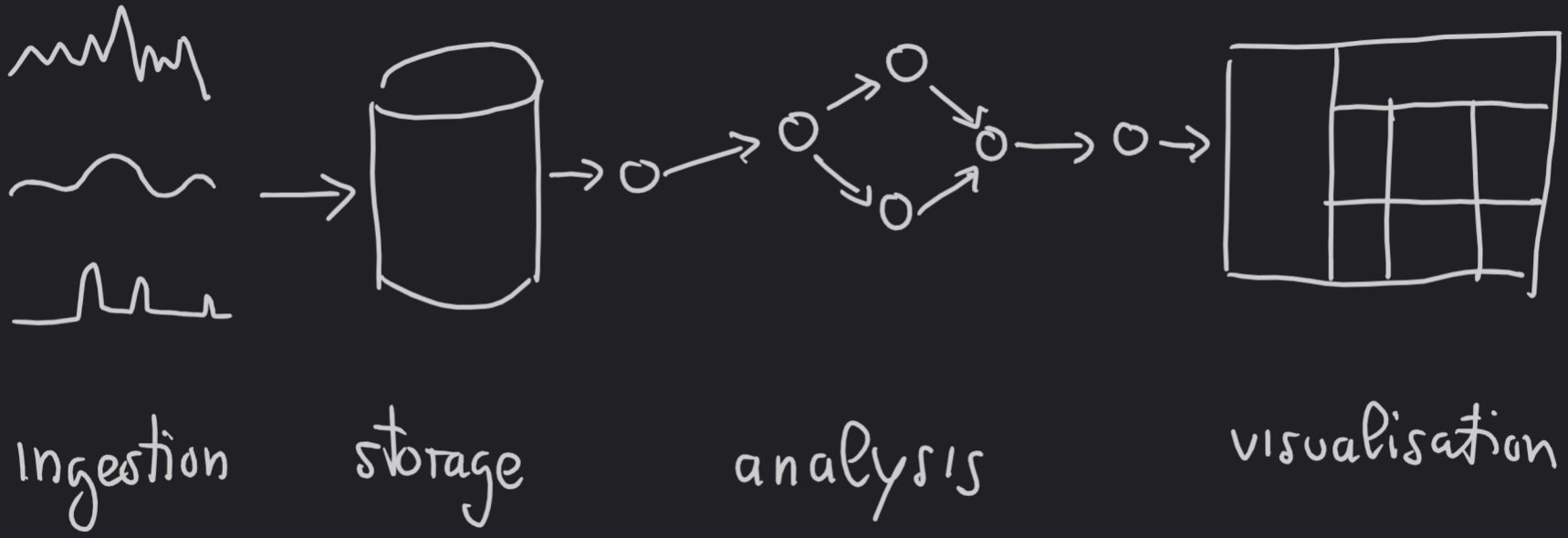
Data Lifecycle



Data Lifecycle

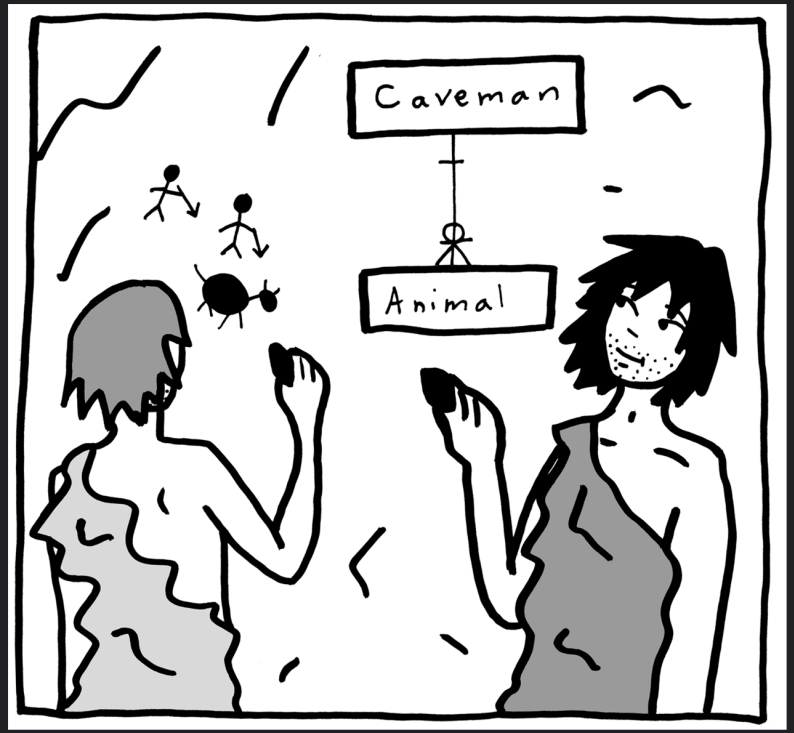


Data Lifecycle



How is data ingested?

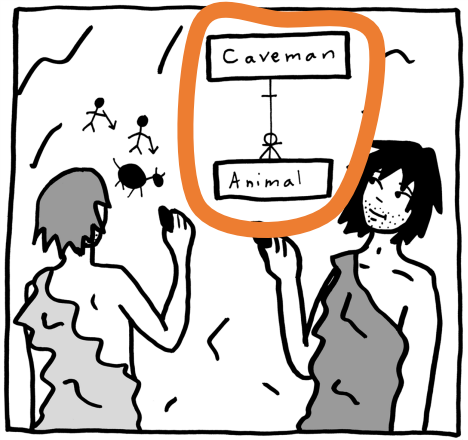
Conceptual vs. logical vs. physical views



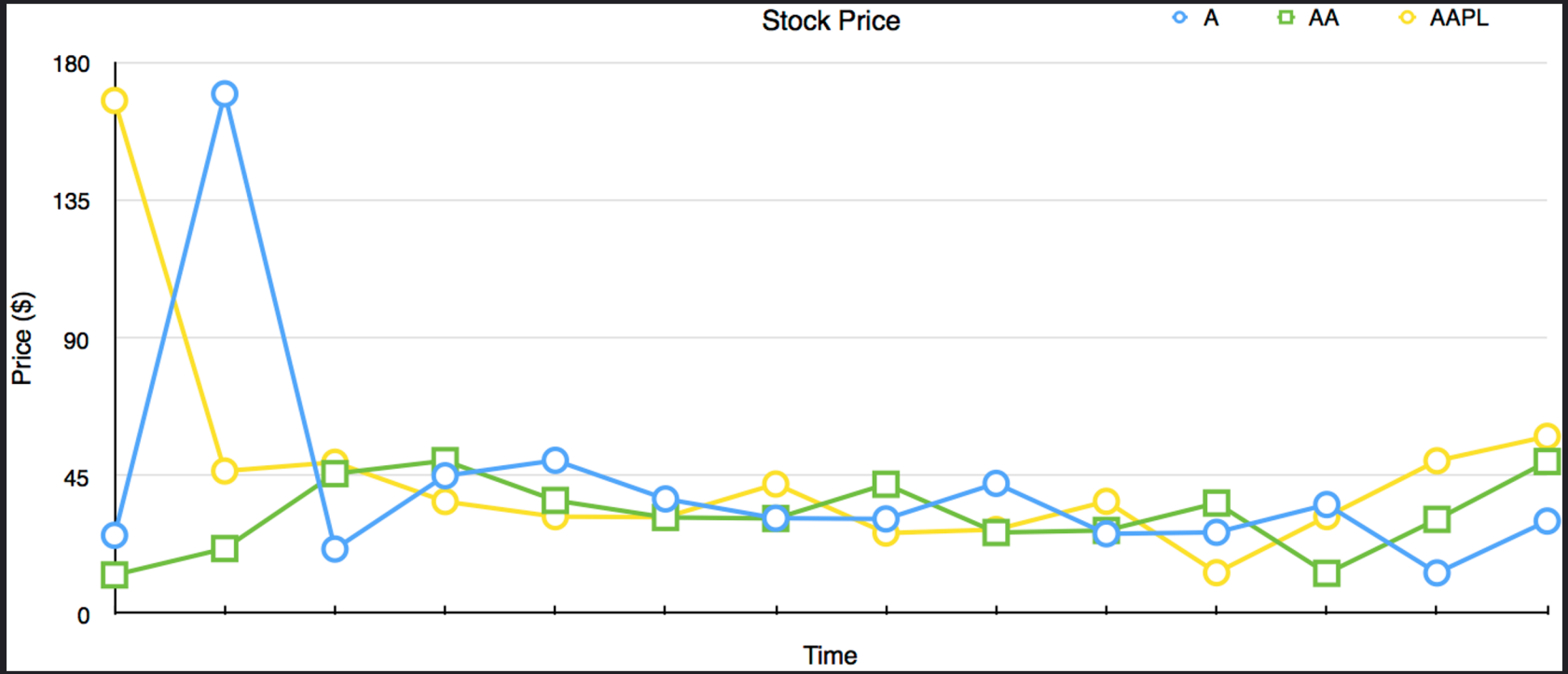
[Src: https://www.oreilly.com/library/view/data-modeling-made/9781935504481/]

How is data ingested?

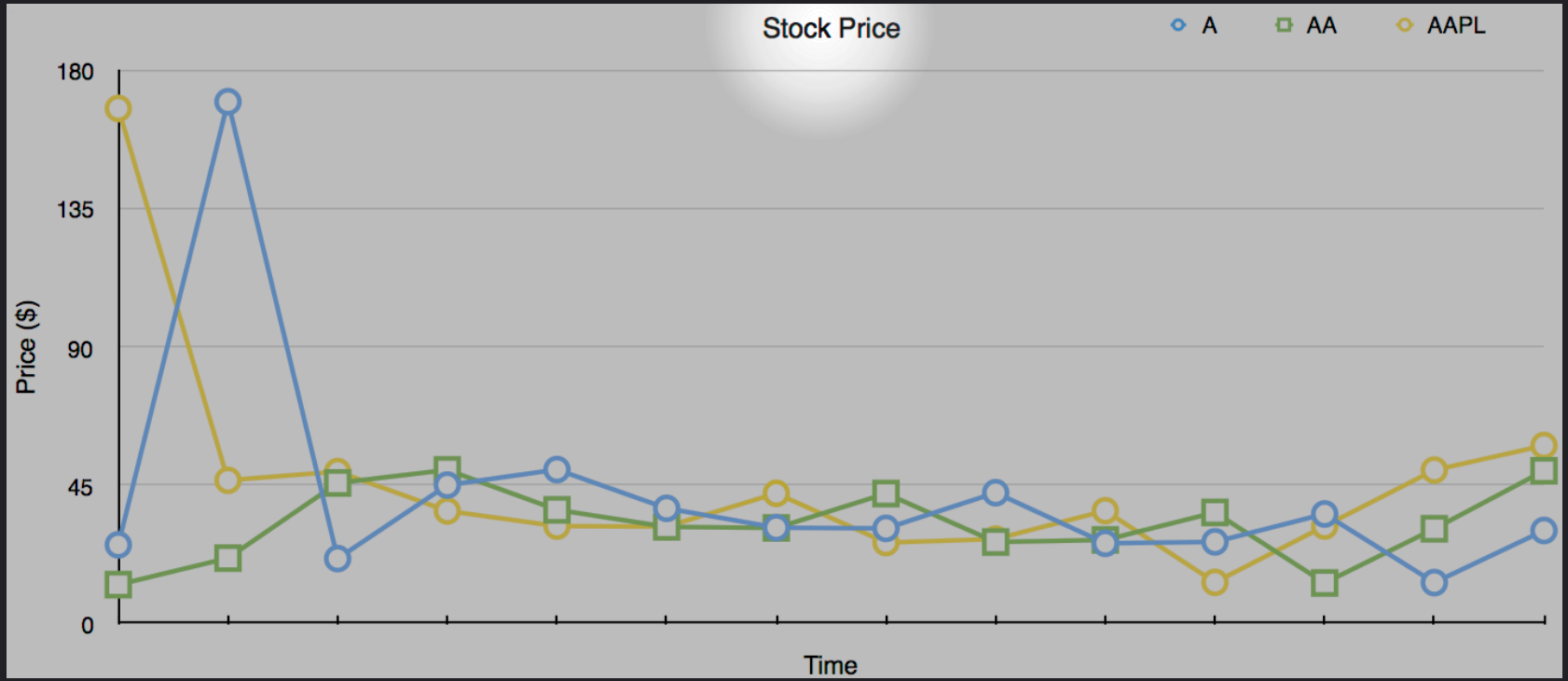
Conceptual View



Let’s start from the anatomy of a Time-Series Line Graph



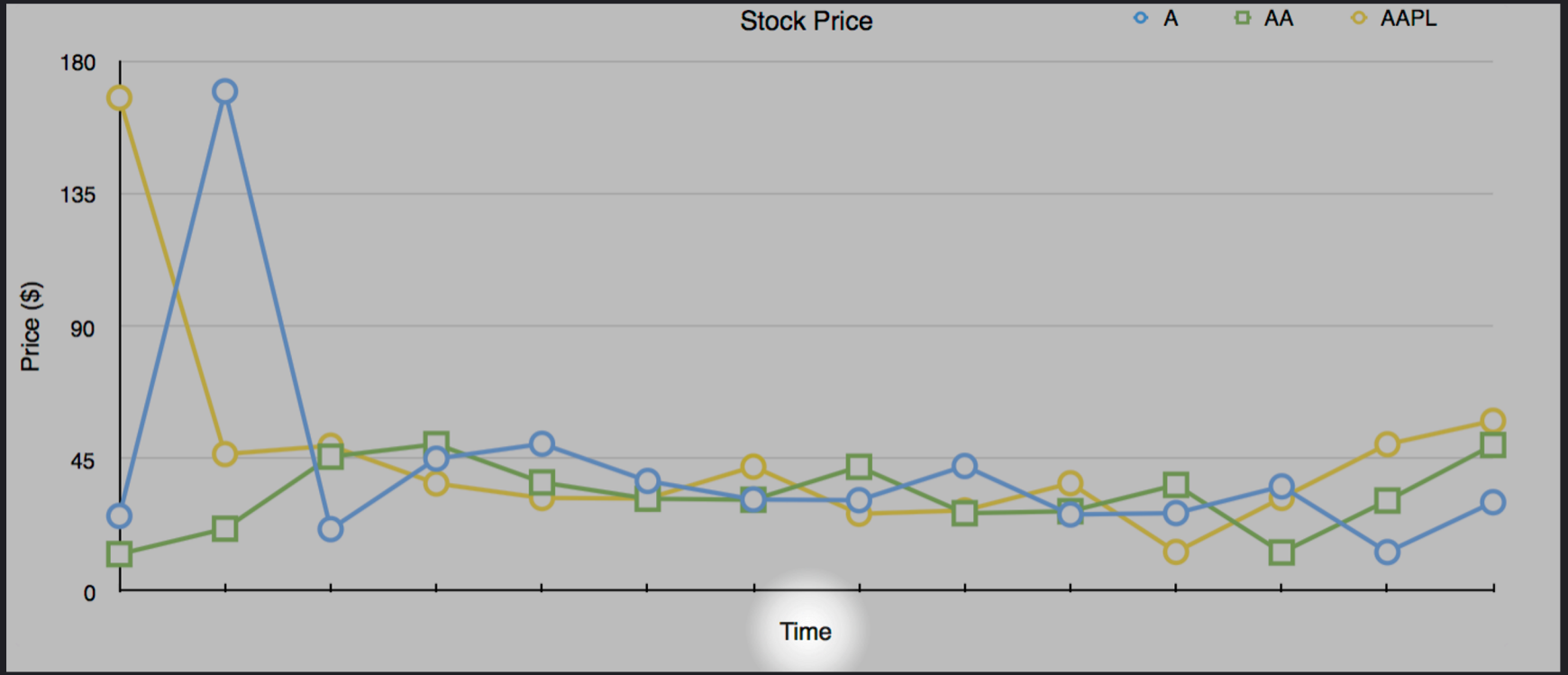
The type of measurement is the title of the Line Graph



Data Model

- Measurement
  - The name of the measurement used as high level grouping of data
-
  -
-
  -
-
  -
-
  -

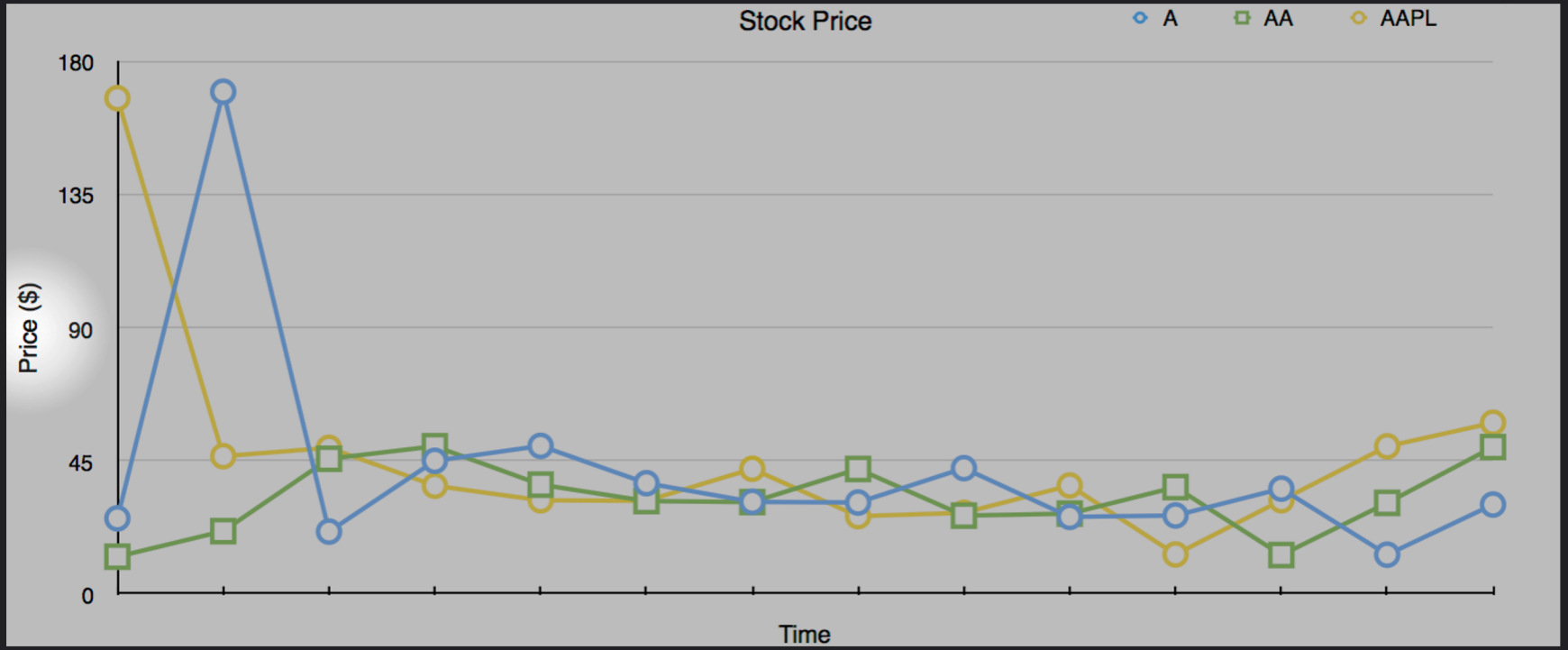
Time stamps are on the X-Axis



Data Model

- Measurement
  - The name of the measurement used as high level grouping of data
-
  -
-
  -
- Timestamp
  - Time of the data (better if the time the data is observed)
-
  -

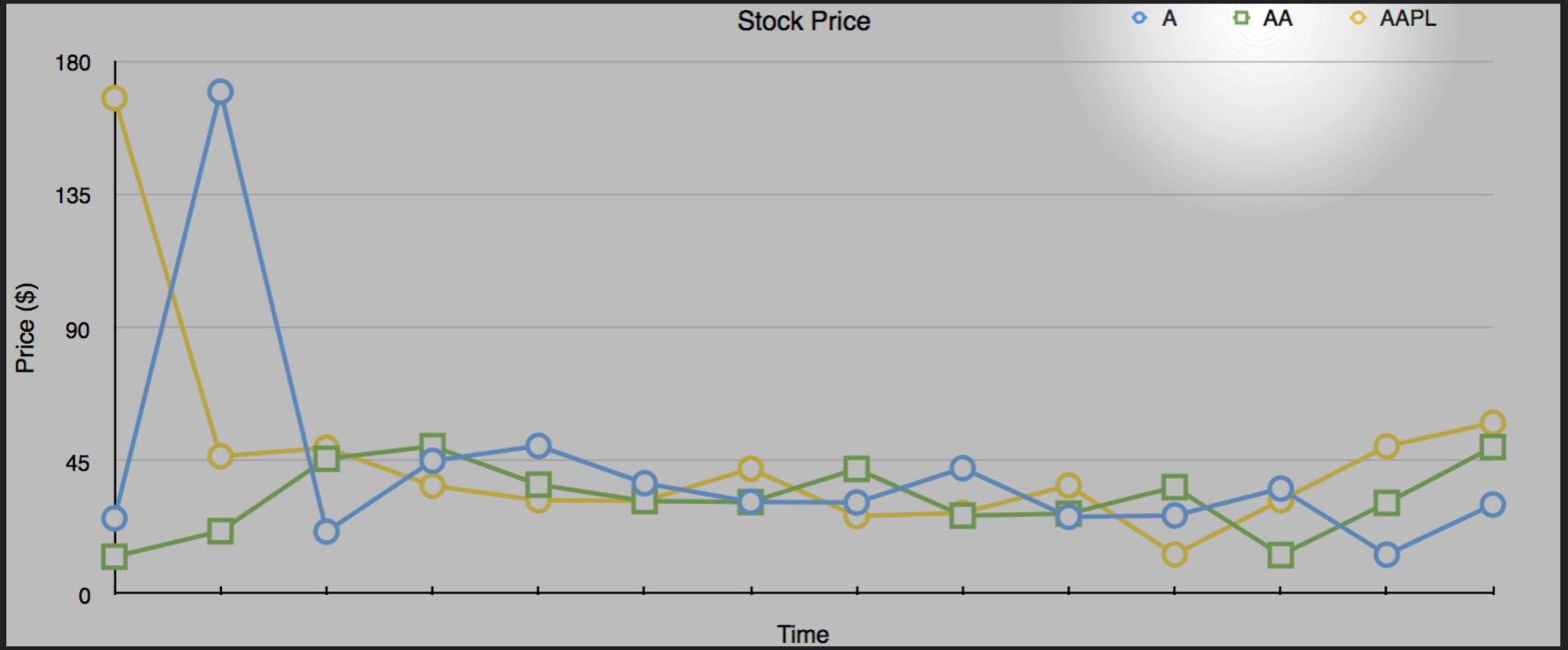
Data is on the Y-Axis



Data Model

- Measurement
  - The name of the measurement used as high level grouping of data
-
  -
- Field set
  - Actual data
- Timestamp
  - Time of the data (better if the time the data is observed)
-
  -

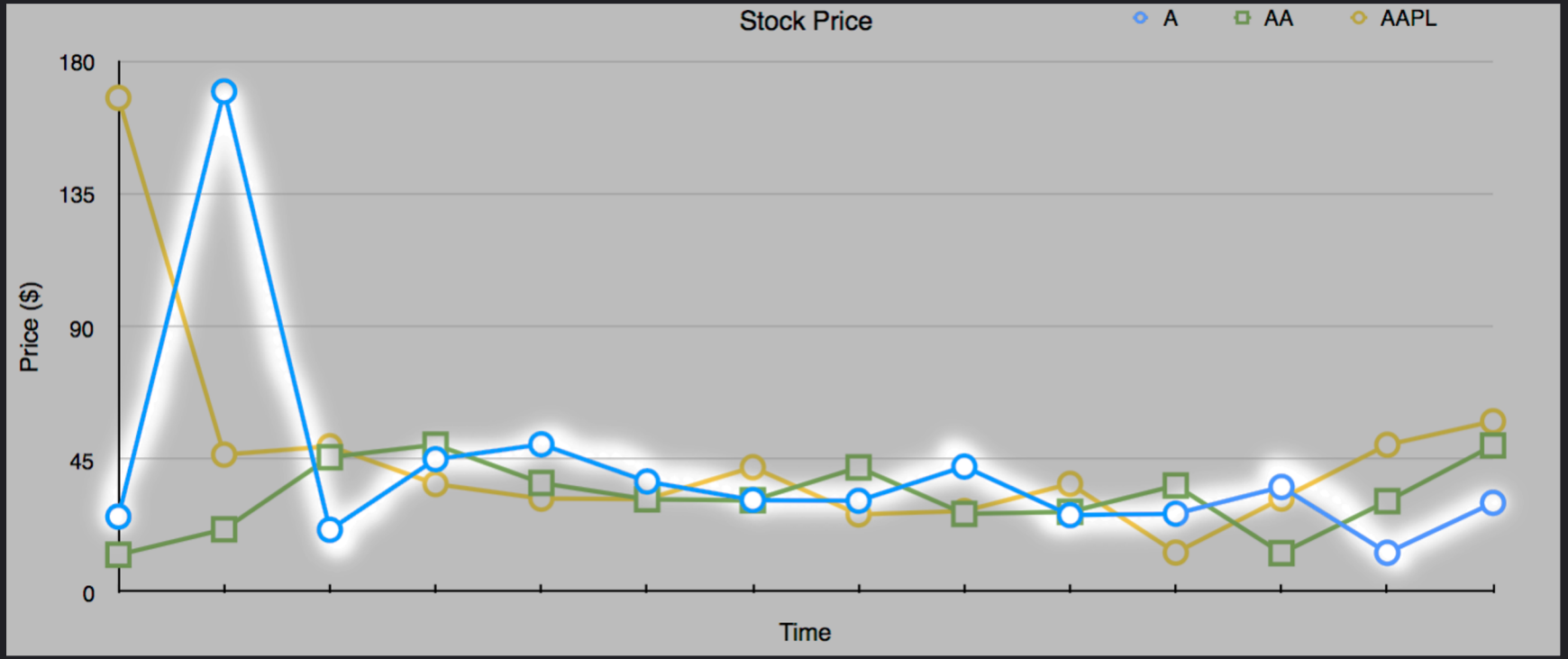
The Legend distinguishes the three time series in the graph



Data Model

- Measurement
  - The name of the measurement used as high level grouping of data
- Tag set
  - Other lower level grouping criteria of data
- Field set
  - Actual data
- Timestamp
  - Time of the data (better if the time the data is observed)
-
  -

A series in the graph thus is …

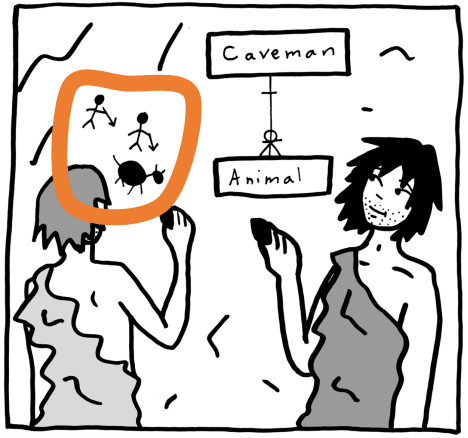


Data Model

- Measurement
  - The name of the measurement used as high level grouping of data
- Tag set
  - Other lower level grouping criteria of data
- Field set
  - Actual data
- Timestamp
  - Time of the data (better if the time the data is observed)
- Series
  - Data points in time order grouped by measurements, tags and fields

How is data ingested?

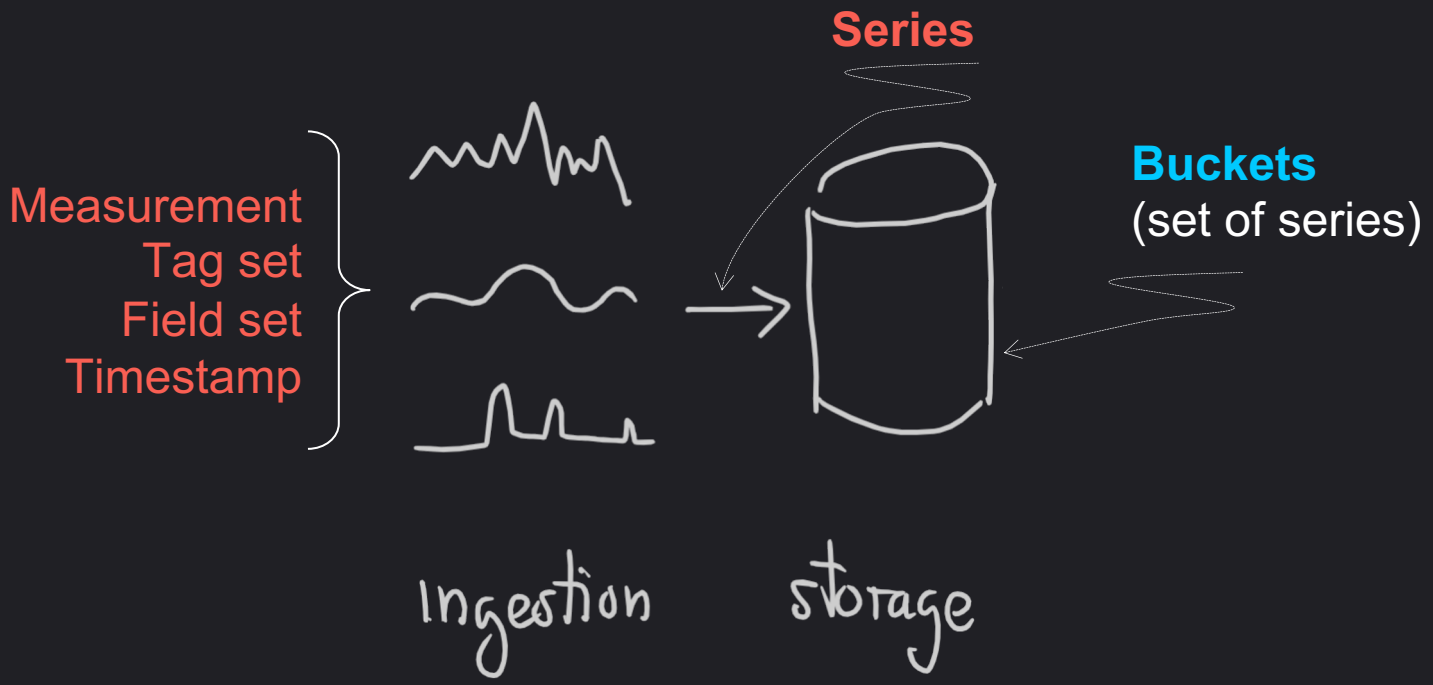
Logical View



Data Model

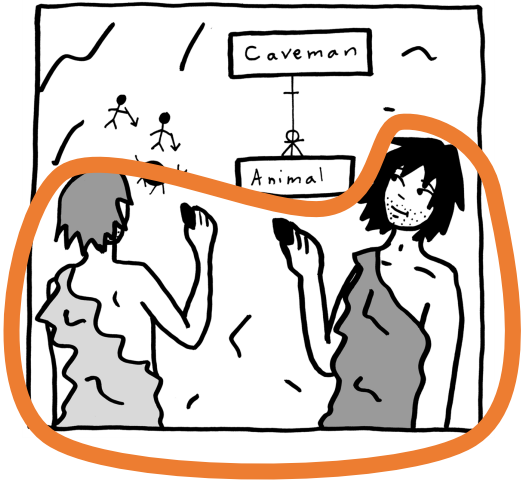
- Measurement
  - A **name** to group data at high level
- Tag set
  - A set of **key-value pairs** to group data at low level
- Field set
  - A set of **key-value pairs** to represent data
- Timestamp
  - **Time** of the data **with nanosecond precision**
- Series
  - A unique combination of measurement+tags+field keys

Data model vs ingestion & storage



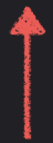
How is data ingested?

Physical View



An example of Line Protocol

cpu,host=serverA,num=1,region=west idle=1.667,system=2342.2 1492214400000000000



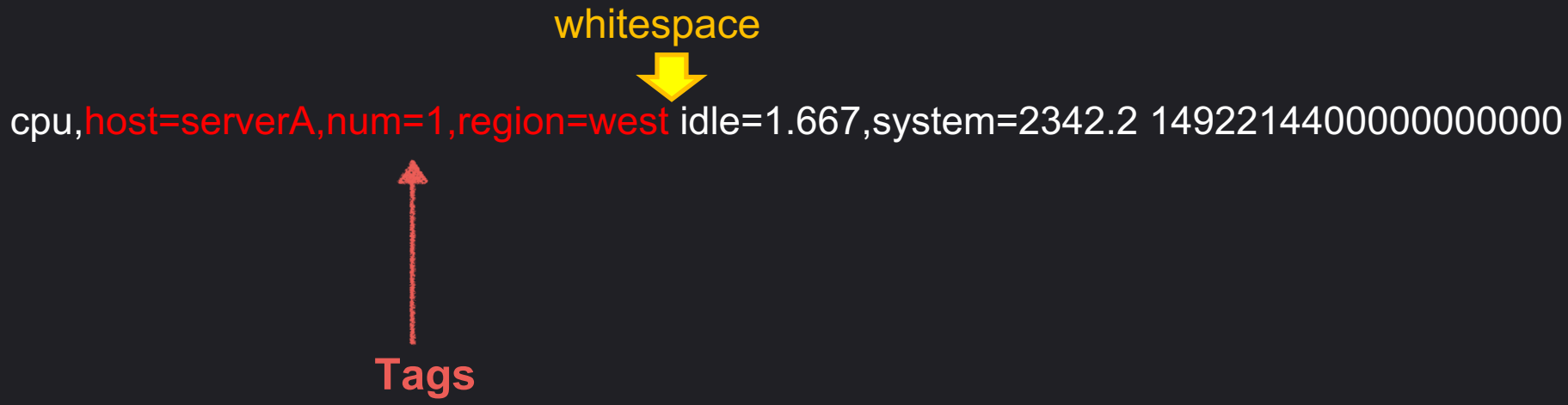
Measurement

An example of Line Protocol

cpu,host=serverA,num=1,region=west idle=1.667,system=2342.2 14922144000000000000

Tags

An example of Line Protocol

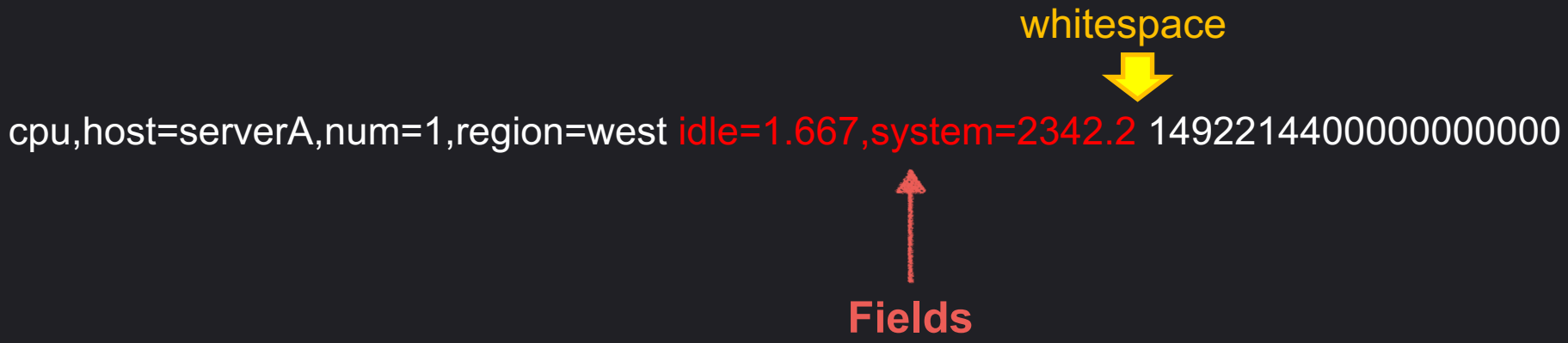


An example of Line Protocol

cpu,host=serverA,num=1,region=west idle=1.667,system=2342.2 14922144000000000000

Fields

An example of Line Protocol



An example of Line Protocol

cpu,host=serverA,num=1,region=west idle=1.667,system=2342.2 1492214400000000000

timestamp

Reference: https://v2.docs.influxdata.com/v2.0/reference/line-protocol/

Bucket physical view

- Columnar Data Stores
- Best for column selection and aggregation thanks to
  - Disk + Memory locality
  - Cache locality

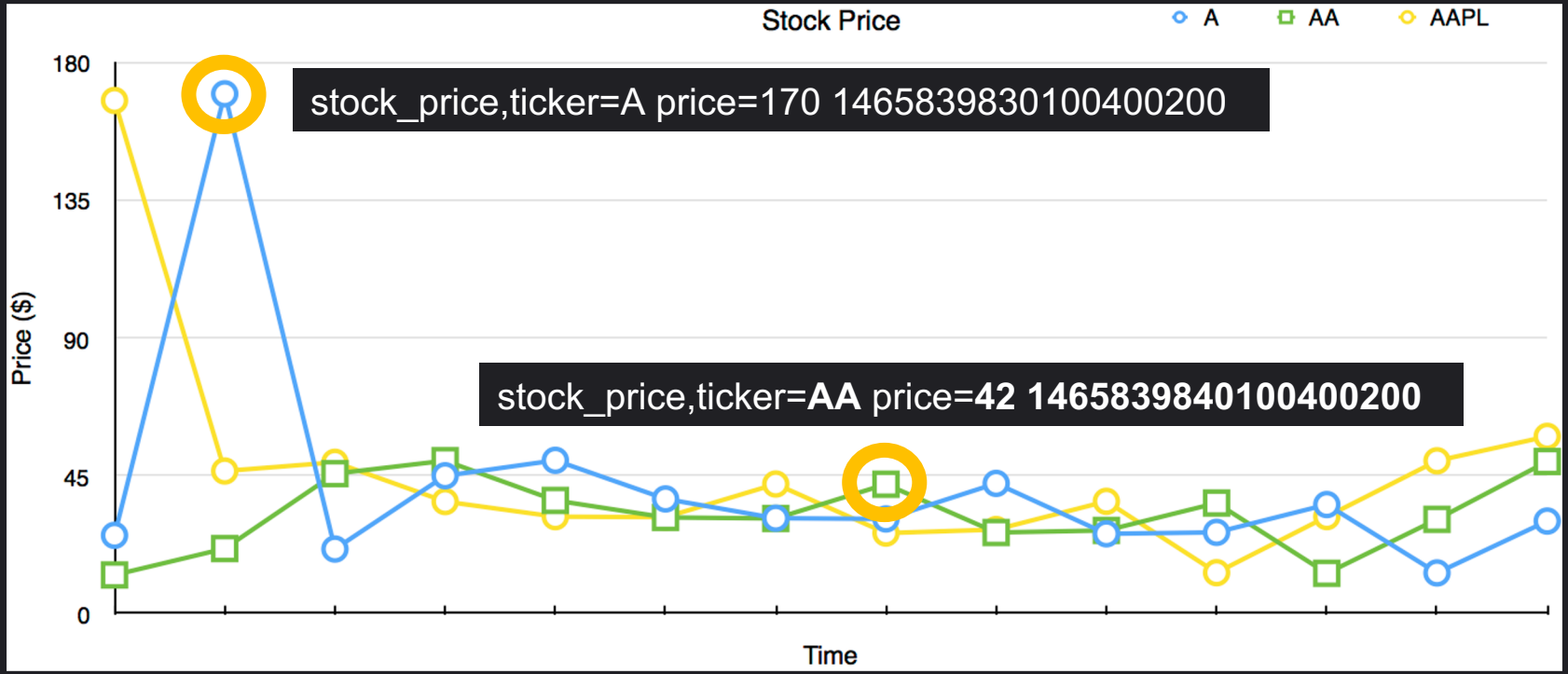
_time	_m	host	num	region	idle	system
1492…1	cpu	serverB	1	west	1.667	2342.2
1492…0	cpu	serverA	1	west	1789.4	2.779
…	…	…	…	…	…	…

tags

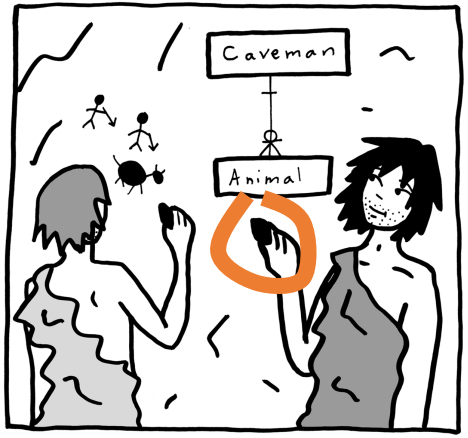
fields

_m = _measurement

So, the line protocol representations of two metrics are …

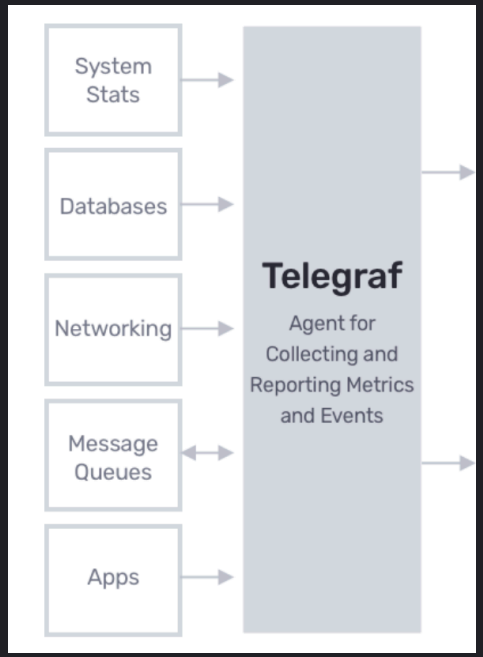


How is data automatically ingested?



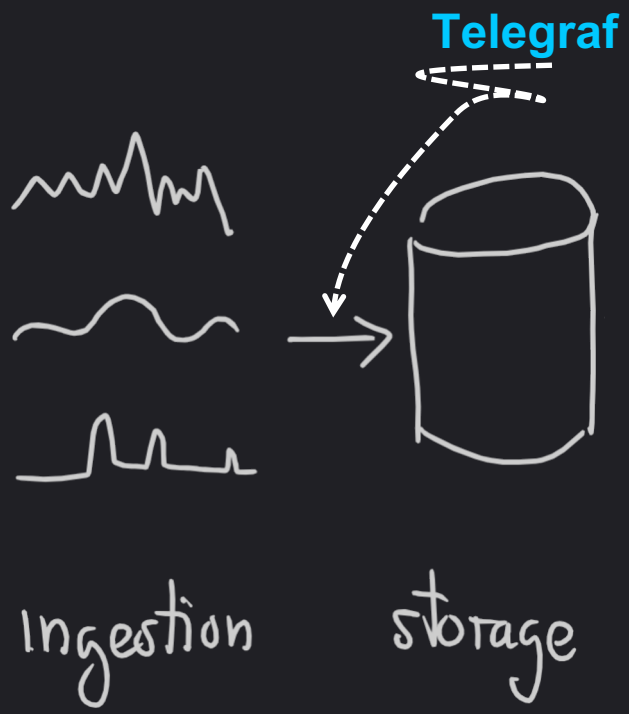
Telegraf

- Telegraf is a data collection agent
- It is based on a Plug and Play architecture
- It offers a variety of input plugins
- It can be configured from the InfluxDB cloud UI



- Download and install: https://docs.influxdata.com/telegraf/latest/introduction/installation/

Positioning Telegraf in the ingestion pipeline



Quiz

- Provide the correct relationship — 1:1, 1:N, N:1, or N:N
  - a time series to a measurement — ?
  - a time series to a tag — ?
  - a time series to a field name — ?
  - a field name to a field value — ?
  - a time series to a timestamp — ?
  - a `<measurement,tag,field name,timestamp>` to a value — ?

Quiz answers

- Provide the correct relationship — 1:1, 1:N, N:1, or N:N –
  - a time series to a measurement — 1:1
  - a time series to a tag — 1:N
  - a time series to a field name — 1:N
  - a field name to a field value — 1:N
  - a time series to a timestamp — 1:N
  - a `<measurement,tag,field name,timestamp>` to a value — 1:1

Let’s get dirty!

1

Use Case: Continuous Linear Pizza Oven

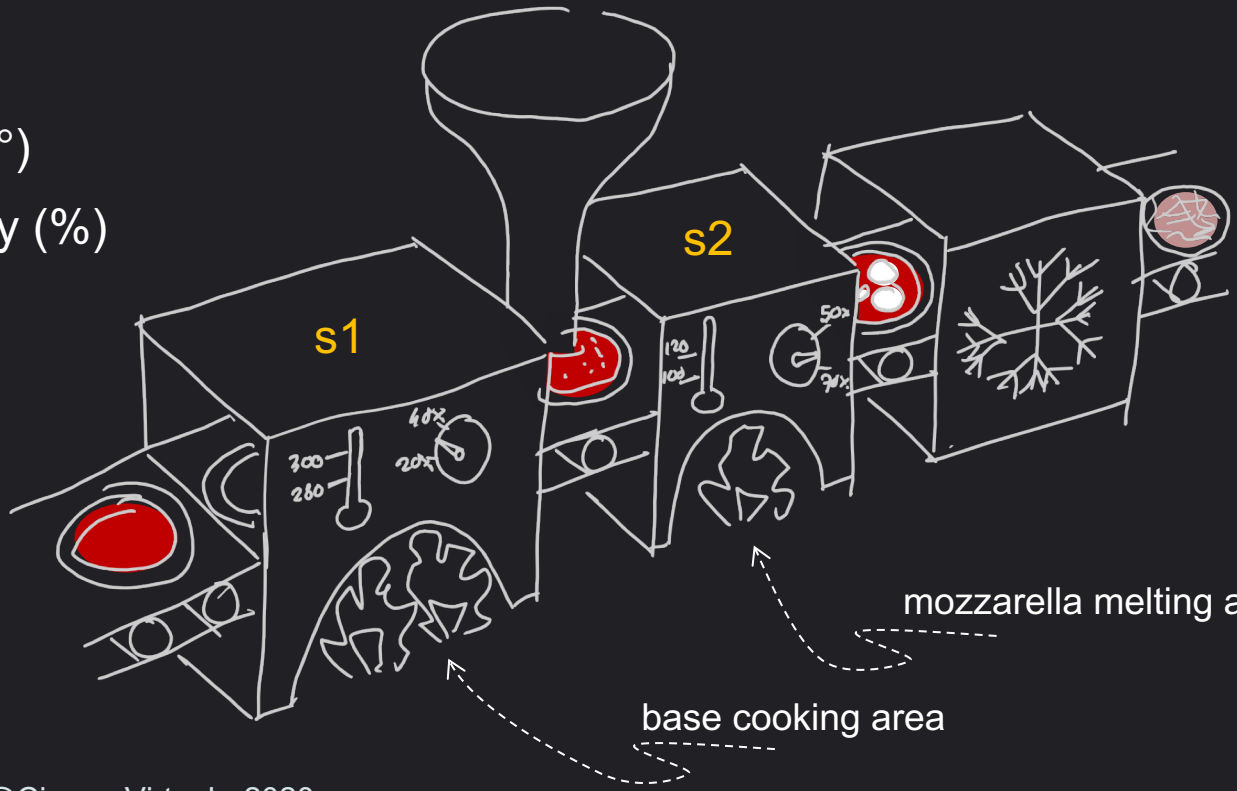
Sensors observe

- temperature (C°)
- relative humidity (%)

of the two ovens

Learning Goals

- Line protocol usage
- First query



Task 1

Model the following data representing the temperature and the humidity observations, from both sensors, over time.

component	sensor	temperature	humidity	ts
iot-oven	S1	290	30	1569888000000000000
iot-oven	S2	105	55	1569888015000000000
iot-oven	S1	305	38	1569888060000000000
iot-oven	S2	120	65	1569888075000000000

Task 2

Run you first query

```
from(bucket: "training")
|> range(start: 2019-10-01T00:00:00Z, stop: 2019-10-01T00:05:00Z)
|> filter(fn: (r) => r._measurement == "iot-oven")
|> filter(fn: (r) => r._field == "temperature")
|> filter(fn: (r) => r.sensor == "S2")
|> filter(fn: (r) => r._value > 100)
```

Quiz

- Should a bucket contain the same number of distinct measurements, tag names, tag values, field names, field values, and timestamps the same?
- If not, can you order them?
  use:
  - >> (much more than)
  - ~ (around the same)

  e.g., field values >> field names

Quiz answers

- Should a bucket contain the same number of distinct measurements, tag names, tag values, field names, field values, and timestamps the same?
- If not, can you order them?
  use:
  - >> (much more than)
  - ~ (around the same)
- ts >> field values >> tag values >> field names ~ tag names >> measurements

Data Ingestion

Emanuele Della Valle Prof. @ Politecnico di Milano & Partner @ Quantia Consulting
Marco Balduini Founder & CEO @ Quantia Consulting