UNIVERSITÀ
DEGLI STUDI
DI PADOVA

J. Pazzini
PADOVA UNIVERSITY, INFN

1 - INTRO TO DATA MANAGEMENT

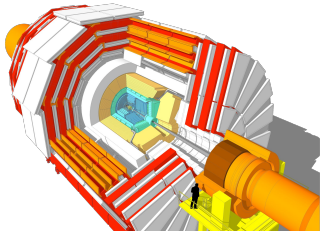
Management and Analysis of Physics Datasets - Module B

Physics of Data

A.A. 2023/2024

FROM DATA TO A DATASET

Raw data

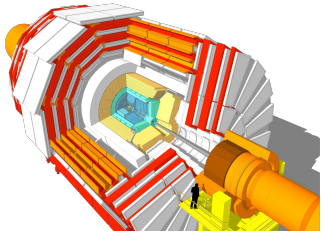


DAQ

Write continuously
at high-rate

FROM DATA TO A DATASET

Raw data



DAQ

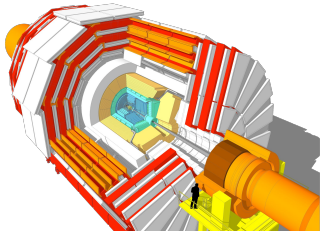
Write continuously at high-rate

Processing

Write every few hours / days

FROM DATA TO A DATASET

Raw data



DAQ

Write continuously at high-rate

Processing

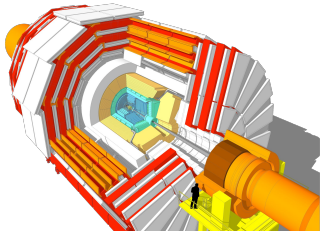
Write every few hours / days

Reprocessing

Write every ~year

FROM DATA TO A DATASET

Raw data



DAQ

Write continuously at high-rate

Processing

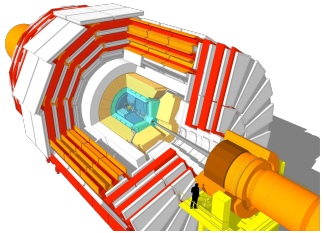
Write every few hours / days

Reprocessing

Write every ~year

\+ **(MC) Simulations**

FROM DATA TO A DATASET



Raw data

\+ **(MC) Simulations**

DAQ

Write continuously at high-rate

Processing

Write every few hours / days

Reprocessing

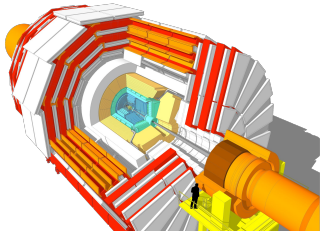
Write every ~year

Long term storage

Write ~once

FROM DATA TO A DATASET

Raw data



+ (MC) Simulations

DAQ

Write continuously at high-rate

Processing

Write every few hours / days

Reprocessing

Write every ~year

Long term storage

Write ~once

Times multiple:

- Data acquisition campaigns
- Years of continuous data taking
- Version of Simulation software
- …

DATASET

Dataset → a collection of data…

A set of information (data, simulation, …) **collected by a number of sources** (detectors, IoT devices, …) **throughout a stretch of time** (years, campaigns, …) **associated to a unique field or purpose**

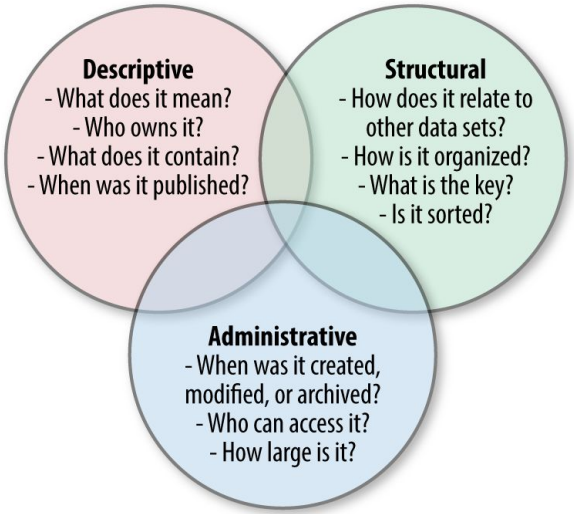
Every single *datum* (“piece of data”) is often referred to as a *record*

DATA ABOUT DATA

Datasets are very often accompanied by *metadata*

Metadata → information about the data itself

They provide valuable information to interpret the data and allow to access/describe/process the dataset correctly



DATA ABOUT DATA

Datasets are very often accompanied by *metadata*

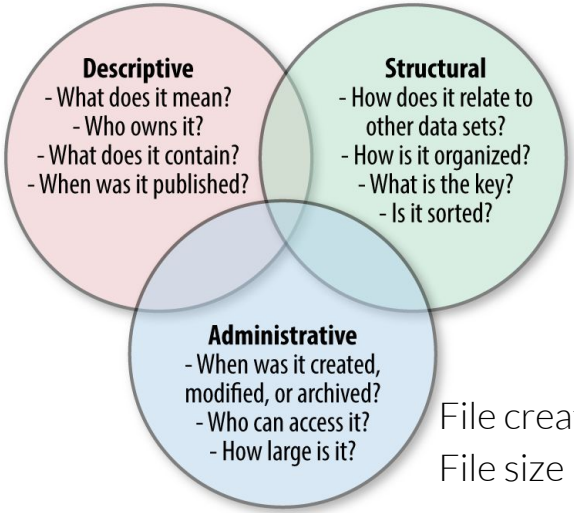
Metadata → information about the data itself

They provide valuable information to interpret the data and allow to access/describe/process the dataset correctly

`a_nice_cat_photo.jpeg`

Caption

Photographer name



Part of the "feline" dataset

File creation/modification date

File size

DATA ABOUT DATA

Datasets are very often accompanied by *metadata*

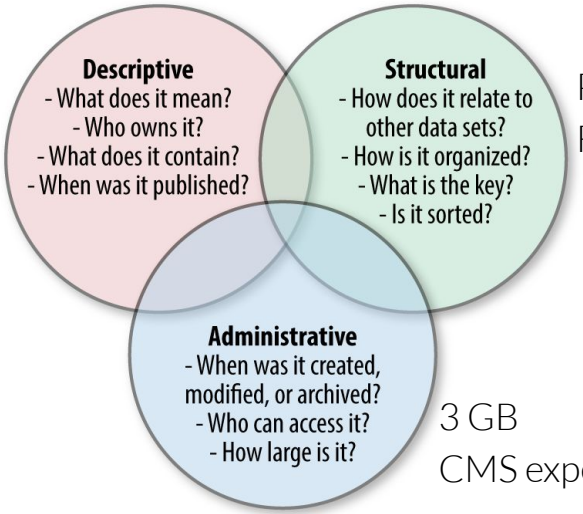
Metadata → information about the data itself

They provide valuable information to interpret the data and allow to access/describe/process the dataset correctly

```
…
0291ce0 3235 3631 3020 3030 3030 6e20 0a20 3030
0291cf0 3230 3836 3235 3933 3020 3030 3030 6e20
0291d00 0a20 3030 3230 3836 3235 3036 3020 3030
0291d10 3030 6e20 0a20 3030 3230 3836 3235 3338
0291d20 3020 3030 3030 6e20 0a20 3030 3230 3836
0291d30 3335 3430 3020 3030 3030 6e20 0a20 3030
0291d40 3230 3836 3335 3632 3020 3030 3030 6e20
0291d50 0a20 3030 3230 3836 3335 3734 3020 3030
0291d60 3030 6e20 0a20 3030 3230 3836 3335 3936
0291d70 3020 3030 3030 6e20 0a20 3030 3230 3836
0291d80 3335 3039 3020 3030 3030 6e20 0a20 3030
0291d90 3230 3836 3435 3331 3020 3030 3030 6e20
…
```

3CBC07C2-4B64-E611-B423-02163E0124E9.root

p-p collision
13 TeV energy



Part of the 2018 dataset
Run-D

3 GB
CMS experiment users only

DATA STRUCTURE

STRUCTURED

SEMI-STRUCTURED

UNSTRUCTURED

Containing a defined data type, format, and structure
→ **data schema known and fixed**

e.g.: .csv, RDBMS* data, …

```
e86d 95d2 0512 407a
4290 95d4 0512 404a
462b 95ed 0512 40c0
10db 95fa 0512 4020
8923 95fa 0512 40f0
c95a 95fe 0512 4068
```

```
1,0,61,42552041,1859,13
1,0,37,42552042,532,16
1,0,96,42552054,2609,11
1,0,16,42552061,134,27
1,0,120,42552061,1097,3
1,0,52,42552063,1610,26
```

HEAD	FPGA	TDC_CHANNEL	ORBIT_CNT	BX_COUNTER	TDC_MEAS
1	0	61	42552041	1859	13
1	0	37	42552042	532	16
1	0	96	42552054	2609	11
1	0	16	42552061	134	27
1	0	120	42552061	1097	3
1	0	52	42552063	1610	26

* Relational DataBase Management System

DATA STRUCTURE

STRUCTURED

SEMI-STRUCTURED

UNSTRUCTURED

Containing a defined data type, format, and structure
→ **data schema known and fixed**

e.g.: .csv, RDBMS data, …

```
e86d 95d2 0512 407a
4290 95d4 0512 404a
461f 65c2 0512 40c0
10db 95fa 0512 4020
8923 95fa 0512 40f0
c95a 95fe 0512 4068
```

```
1,0,61,42552041,1859,13
1,0,37,42552042,532,16
1,0,42552054,96,2609,11
1,0,16,42552061,134,27
1,0,120,42552061,1097,3
1,0,52,42552063,1610,26
```

HEAD	FPGA	TDC_CHANNEL	ORBIT_CNT	BX_COUNTER	TDC_MEAS
1	0	61	42552041	1859	13
1	0	37	42552042	532	16
1	**0**	**42552054**	**96**	**2609**	**11**
1	0	16	42552061	134	27
1	0	120	42552061	1097	3
1	0	52	42552063	1610	26

DATA STRUCTURE

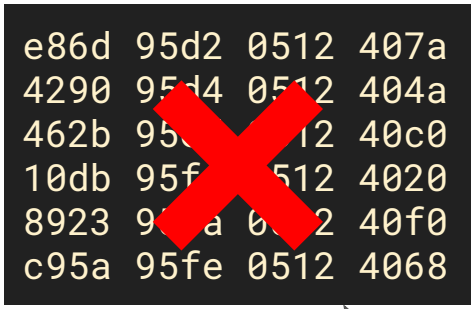
STRUCTURED

SEMI-STRUCTURED

UNSTRUCTURED

Containing a defined data type, format, and structure
→ **data schema known and fixed**

e.g.: .csv, RDBMS data, …



HEAD	FPGA	TDC_CHANNEL	ORBIT_CNT	BX_COUNTER	TDC_MEAS
1	0	61	42552041	1859	13
1	0	37	42552042	532	16
1	0	96	42552054	2609	11
1	0	16	42552061	134	27
1	0	120	42552061	1097	3
1	0	52	42552063	1610	26

NEW_FEAT
12
5

DATA STRUCTURE

STRUCTURED

SEMI-STRUCTURED

UNSTRUCTURED

Data type with format and structured that can be extracted by parsing

→ data schema not necessarily known or fixed a-priori, but inferred from data

e.g.: .xml, .json, … (.yaml, “email”)

eXtensible Markup Language

```
<person>
  <firstName>John</firstName>
  <lastName>Smith</lastName>
  <age>25</age>
  <address>
    <streetAddress>21 2nd Street</streetAddress>
    <city>New York</city>
    <state>NY</state>
    <postalCode>10021</postalCode>
  </address>
  <phoneNumbers>
    <phoneNumber>
      <type>home</type>
      <number>212 555-1234</number>
    </phoneNumber>
  </phoneNumbers>
  <sex>
    <type>male</type>
  </sex>
</person>
```

DATA STRUCTURE

STRUCTURED

SEMI-STRUCTURED

UNSTRUCTURED

Data type with format and structured that can be extracted by parsing

→ data schema not necessarily known or fixed a-priori, but inferred from data

e.g.: .xml, .json, … (.yaml, “email”)

JavaScript Object Notation

```
{
  "first name": "John",
  "last name": "Smith",
  "age": 25,
  "address": {
    "street address": "21 2nd Street",
    "city": "New York",
    "state": "NY",
    "postal code": "10021"
  },
  "phone numbers": [
    {
      "type": "home",
      "number": "212 555-1234"
    },
  ],
  "sex": {
    "type": "male"
  }
}
```

DATA STRUCTURE

STRUCTURED

SEMI-STRUCTURED

UNSTRUCTURED

Data type with format and structured that can be extracted by parsing
→ data schema not necessarily known or fixed a-priori, but inferred from data

e.g.: .xml, .json, … (.yaml, "email")

A schema can also be defined for semi-structured data to:

- Check the integrity / validate data
- Recast the data to structured datasets

First name	Last name	Age	Address	Phone numbers	Sex
John	Smith	25			M

Street address	City	State	Postal code
21 2nd Street	New York	NY	10021

Type	Number
Home	212 555-1234

DATA STRUCTURE

STRUCTURED

SEMI-STRUCTURED

UNSTRUCTURED

Data type with no predefined structure
→ **no data schema available a-priori, nor inferred from data**

Typically text-heavy, plus additional formats
(audio/video/geolocation/dates/numbers/…)

Not suitable for RDBMS, or any schema-driven storage/analytics

DATA STRUCTURE

STRUCTURED

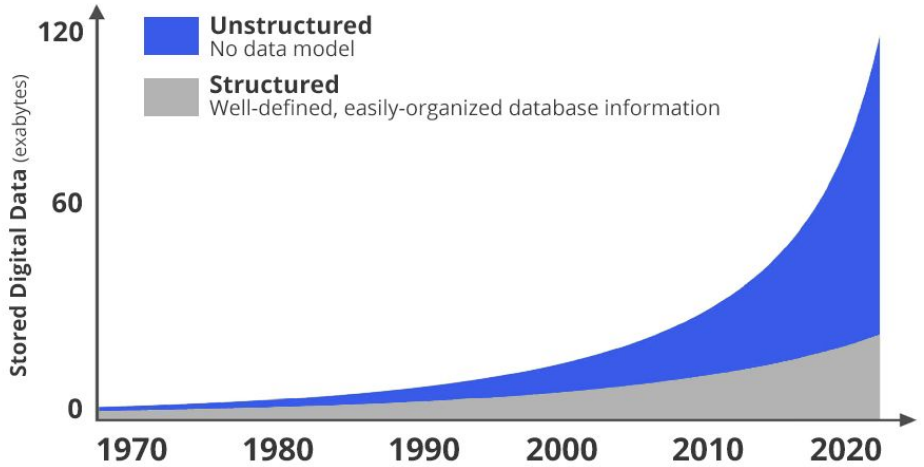
SEMI-STRUCTURED

UNSTRUCTURED

Data type with no predefined structure
→ **no data schema available a-priori, nor inferred from data**

Typically text-heavy, plus additional formats
(audio/video/geolocation/dates/numbers/…)

Not suitable for RDBMS, or any schema-driven storage/analytics



DATA MANAGEMENT

Data Management deals with all the needs & challenges related to the (safe)keeping and accessibility of datasets:

- Data storage
- Reliability & Long-term preservation
- Accessibility & security
- File/Database management systems
- …