1. fundamentals
2. cleaning
3. visual comm
4. comm with data
5. storage and privacy

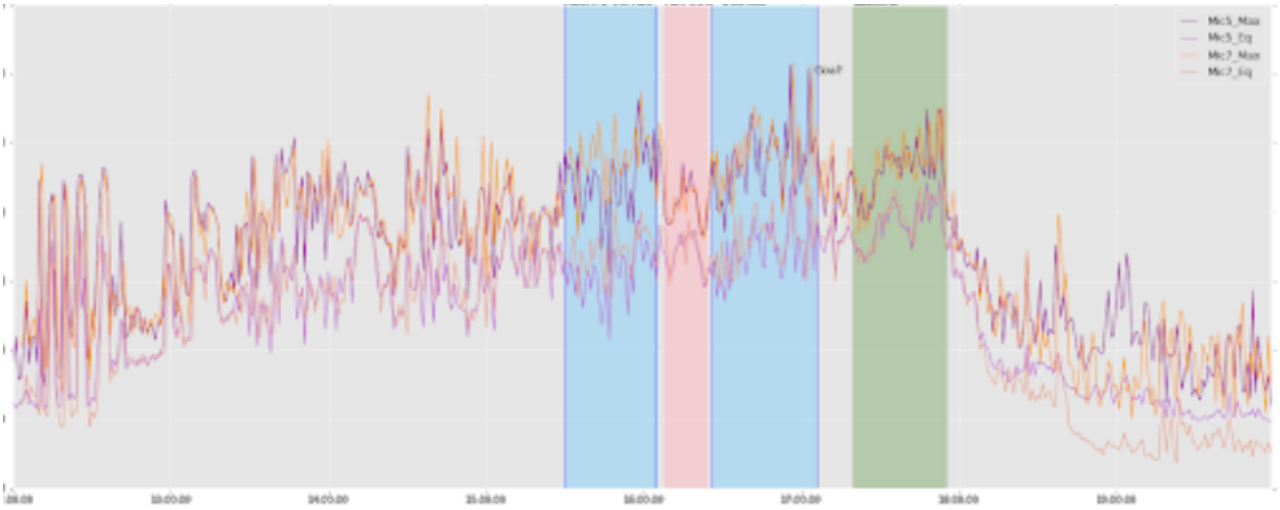
w1:

Data Is The New Oil

data analytics pipeline

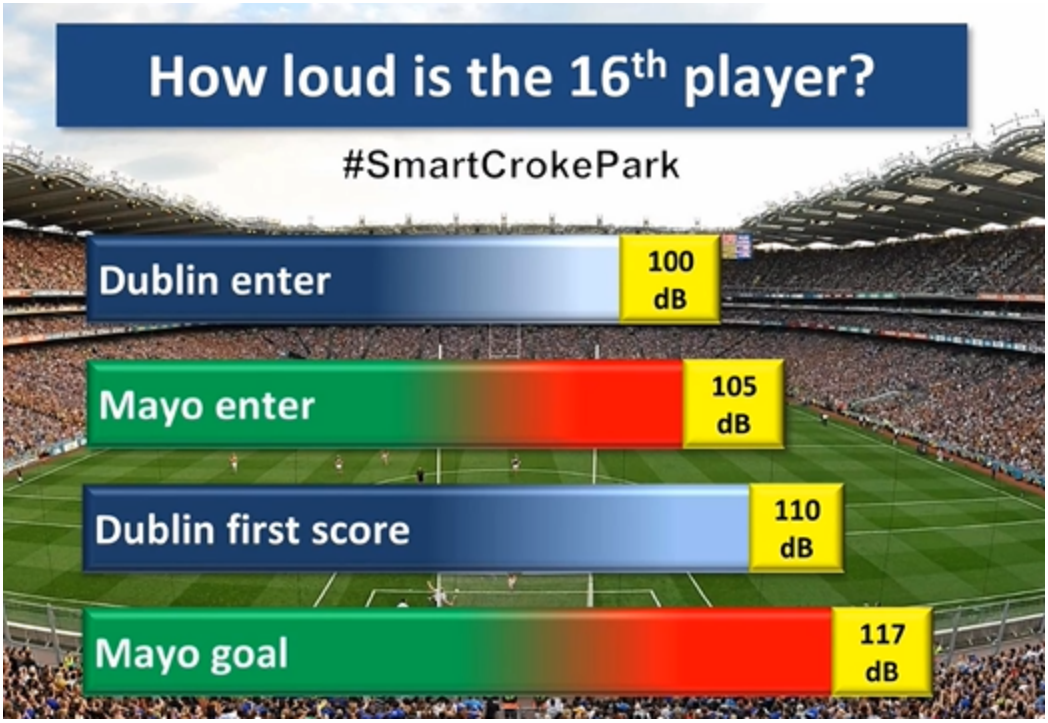
smart stadium

- how load is the 16th player, micro for fan engagement (only measure loudness)
- what actions on field, so visualise on
- was there any decernable dif? we use **explorative visual** (drop at half time)



- spikes cor with actions on pitch (goal) or disagreement on pitch
- 15 mins sound decreases

explanitive vis:



> Exploratory Visualisations help the viewer (usually the analyst) to understand a data set, what data is available, what patterns might be present and if there could be errors or anomalies in the data.
>
> Explanatory Visualisations are designed to communicate a message or idea. The audience should be able to view the visualisation and understand a concept, relationship or one or more facts from the data
>
> Data Drived application: An application that uses data to understand and create value.

w1 quiz: https://www.futurelearn.com/courses/data-management-and-visualisation-data-fundamentals/3/steps/1600092/questions/1

---

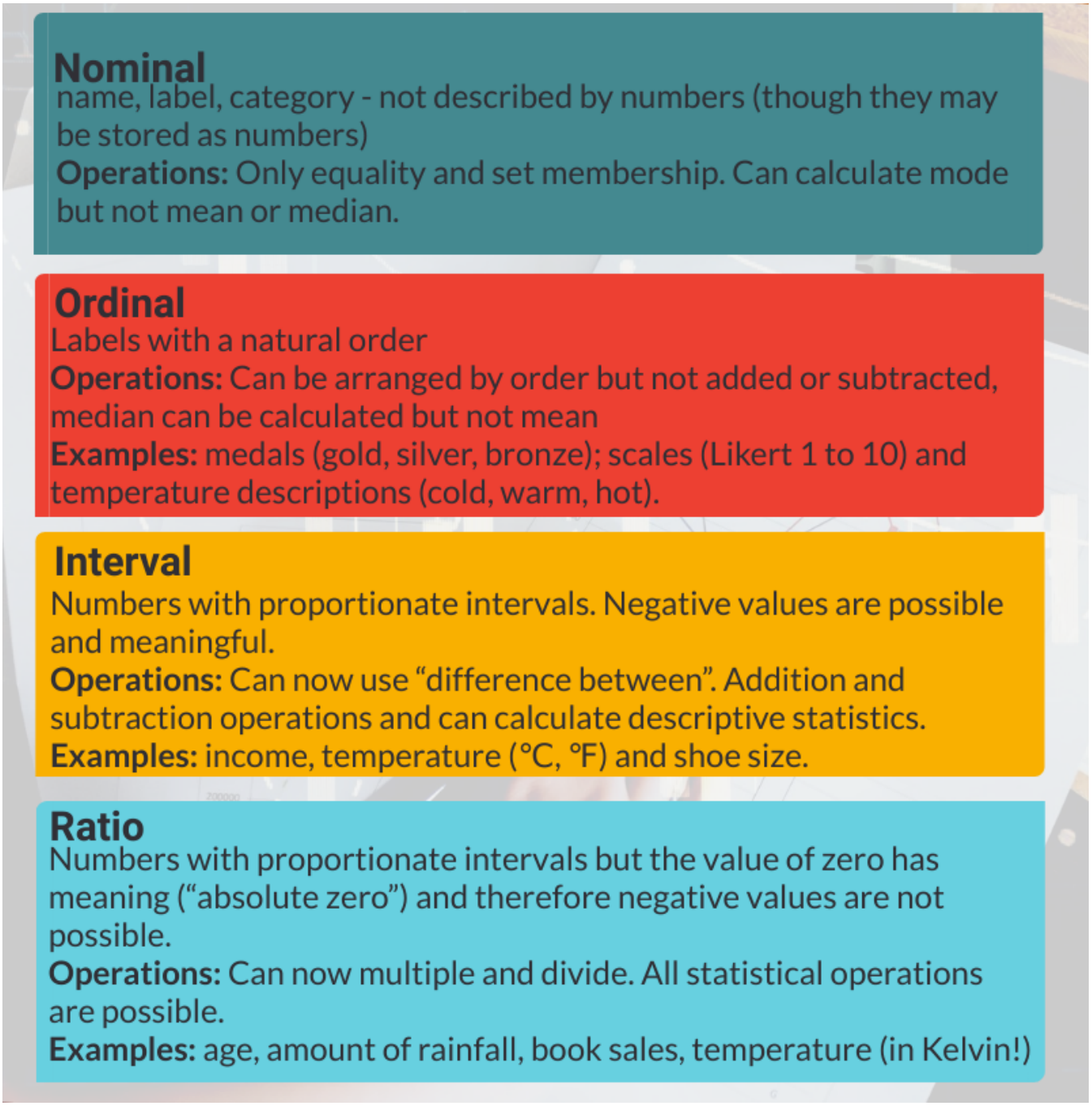
w2

Data types, four different ways of categorising data features:

- Structure (spreadsheet) vs unstructured (images, word doc)
- Quantitative (interval, ratio) vs qualitative (nominal, ordinal)
- Discrete (cant be divided res be 100.5) vs continuous (can be 100.5)
- NOIR levels of data https://www.futurelearn.com/courses/data-management-and-visualisation-data-fundamentals/3/steps/1600099

NOIR:

The Nominal, Ordinal, Interval and Ratio classes of data measurement



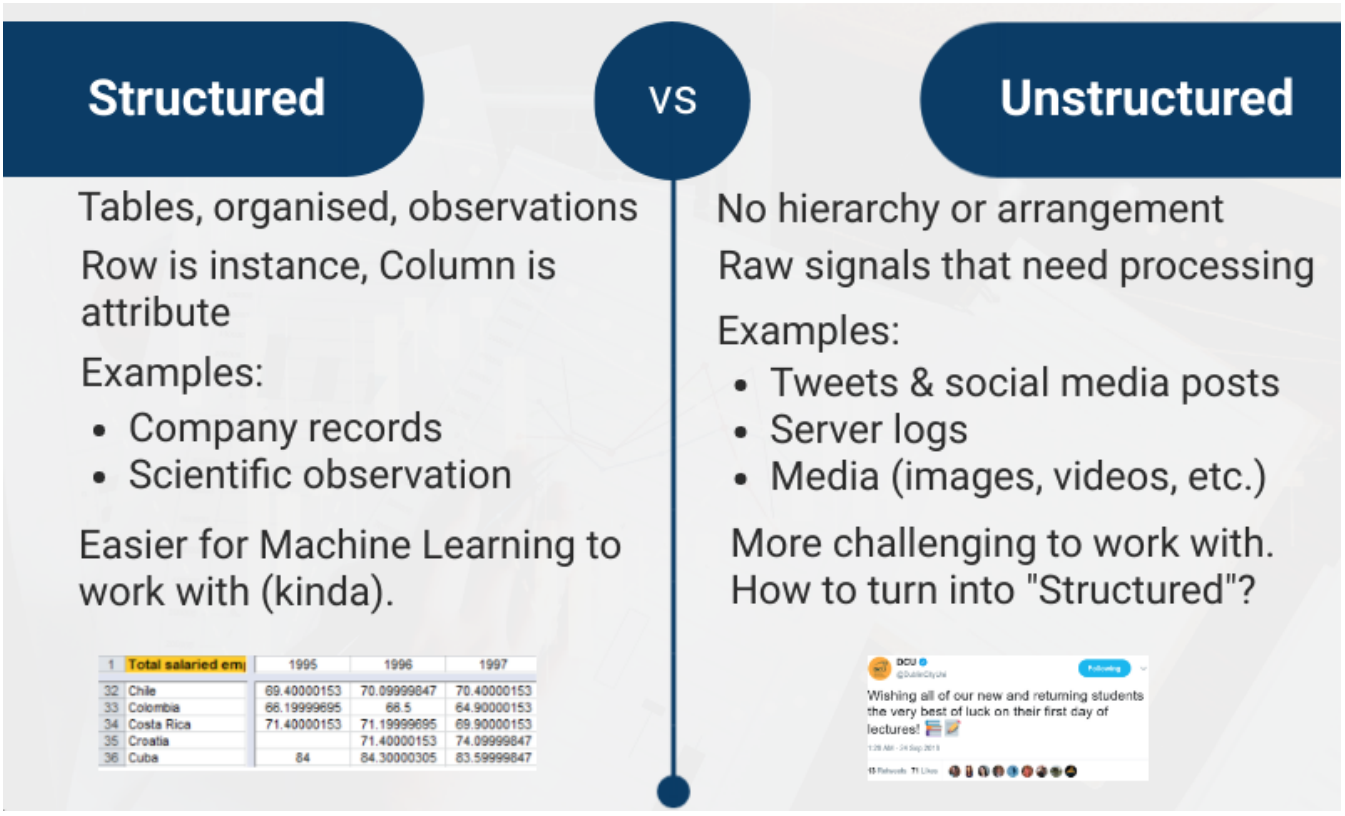
Applying NOIR classification to measurement:

> Temperature is an interesting case. When you use terms to describe the weather (cold, cool, warm, hot) then these are labels with a natural order so this measurement is ordinal. If you measure temperature using a standard system like Celcius or Farenheit then it is an interval measurement because the difference between 10 & 20 and 20 & 30 has the same meaning (the interval is equivalent). It's also possible to have negative degrees Celcius and Fahrenheit so a value of zero does not mean that there is no temperature. However, if you are using the Kelvin scale (for example, in physics and chemistry) then zero is a possible and meaningful value that means there is no physical heat! This becomes a ratio measurement.

other special measurements:

- Time series
- GeoSpatial
  - Calculating the distance between spatial positions or plotting the route or trajectory may help to reveal extra information.
- Documents
- Multimedia

wk2 quiz: https://www.futurelearn.com/courses/data-management-and-visualisation-data-fundamentals/3/steps/1600101/quiz/introduction



Data formats

- Text or binary.
- Open or proprietary.
- Structured or unstructured.

Proprietary formats belong to a company and usually require a specific programme to read and manipulate the contents.

The most common structured or tabulated data format is CSV comma separated values.

Other text-based structured data formats that you should be familiar with include: JSON, XML, HTML

Many database programs (such as SQLite, MariaDB and PostgreSQL) may use text files to store the data, e.g., DB, SQL. The programs are optimised to process, store and retrieve from these files but often they are text-based and can be viewed and read by other programs.

There are other specialist data formats (for example, GDP and ASX) some of which may require a specific programme or library to read.

Databases:

- Relational DB, tables (relations) of rows and col with primary keys to ref between them
- NoSQL db , document-based storage
- in memory database (SAP Hana) for fast resp time

When you are accessing data from a database you will generally use an API, sometimes with a query language

---

wk3

Big Data

Big data is typically characterised by the three “V”s (Volume, Velocity and Variety).

A common fourth V (Veracity) is sometimes added

1. Big data: Volume

Referes to the *amount* of data

1. Big data: Variety

Refers to differing types and data sources (structured, semi-structured, unstructured)

Organisations may need to combine data from many different sources

1. Big data: Velocity

Data in *motion*(dynamic, temporal, streaming).

We may need to process data as it arrives (i.e., time-sensitive, cant have high latency)

1. Big data: Veracity

the tendency of big data to be *unreliable*.

with the huge volumes of data being generated errors will slip in

---

examples of big data:

see internet minute https://www.allaccess.com/merge/archive/31294/infographic-what-happens-in-an-internet-minute

- internet a source of data
- YouTube has 500 hours of video uploaded per minute
- transport a source of data (advanced driver assistance)
- Financial transactions, such as those on the New York Stock Exchange

Big data issues:

- High volume data, system can crash when reached working memory limits
- High velocity data, programmes need to be able to handle a dynamic data stack and should process the new information fast enough to be able to use the conclusions.
- Also, high variety of big data creates some practical challenges

There are ways to **work around** the data limits of standard programmes!

1. working in batches
2. filtering unneeded data entries
3. using a different application or purchasing more memory

Tools for big data:

We'll look briefly at some of the options for storing and processing data (like MapReduce, Spark and Hadoop) in course 5

Apache

Apache manages a number of open-source projects that includes frameworks and tools for specific data ingestion, processing, analytics.

Apache Hadoop provides "a framework that allows for the distributed processing of large data sets across clusters of computers using simple programming models". Apache Hadoop consists of five core modules: Common Utilities, the Hadoop Distributed File System (HDFS) for storage, YARN for process scheduling, MapReduce for parallel processing and Ozone for object storage. Other common projects that operate with the Hadoop stack include:

Apache Spark: "fast and general compute engine for Hadoop data", an alternative to MapReduce, Spark works with different datastores (batch, real-time) and performs in-memory data processing.

Apache Cassandra: "A scalable multi-master database with no single points of failure."

Apache Storm is "a free and open-source distributed realtime computation system". Storm works with unbounded streams of data, doing for realtime processing what Hadoop did for batch processing. You can access a Comparison of Hadoop and Storm here.

Apache Flink: “a framework and distributed processing engine for stateful computations over unbounded and bounded data streams”

Databases or Data access

MongoDB: “a general-purpose, document-based, distributed database”. MongoDB is a cross-platform, open-source, NoSQL database that stores data in JSON-like documents.

Apache CouchDB: JSON document storage, queryable via JavaScript, that has distributed scaling with fault-tolerant storage.

Neo4j: a graph database focussing on data and its relationships at speed and supporting graph algorithms. See also dgraph for an alternative graph database.

ElasticStack (ELK): “ELK” is the acronym for three open-source projects. Elasticsearch is a search and analytics engine. Logstash is a server-side data processing pipeline. Kibana lets users visualize data. ElasticStack has expanded to include other functionality so ELK was renamed.

Data Analytics Platforms

RapidMiner: a software platform for data science activities, this tool is more focused on the analytics phase but allows visual exploration of the output of data-driven models. It also has a GUI for designing workflows while the backend can operate in the cloud increasing processing and storage capacity.

Pentaho: “data integration and analytics platform [that] enables organizations to access, prepare, and analyze all data from any source, in any environment.”

Data Cleaning and Exploration

OpenRefine: open-source application to work with messy data, cleaning it and transforming it from one format into another and extending it with web services and external data.

Tableau for big data: Tableau is a visual analytics platform, Tableau Prep lets you combine, shape, and clean your data. We will have some tasks using Tableau for visualisation in the fourth course in this program. Learn more about how Tableau handles big data here.

Non-tabular data: linked data

"did you mean" recommendations

linked open data, lots if indiv websites as Raw data because it can be linked to other data (creates a network)

1. data open license
2. data on web in certain way, resource description RDF (can be recorded as txt doc in XML format, forms the links.
3. unique URI address
4. HTTP protocol

when other websites link to your data, gets more visible (search engines give it higher weight)

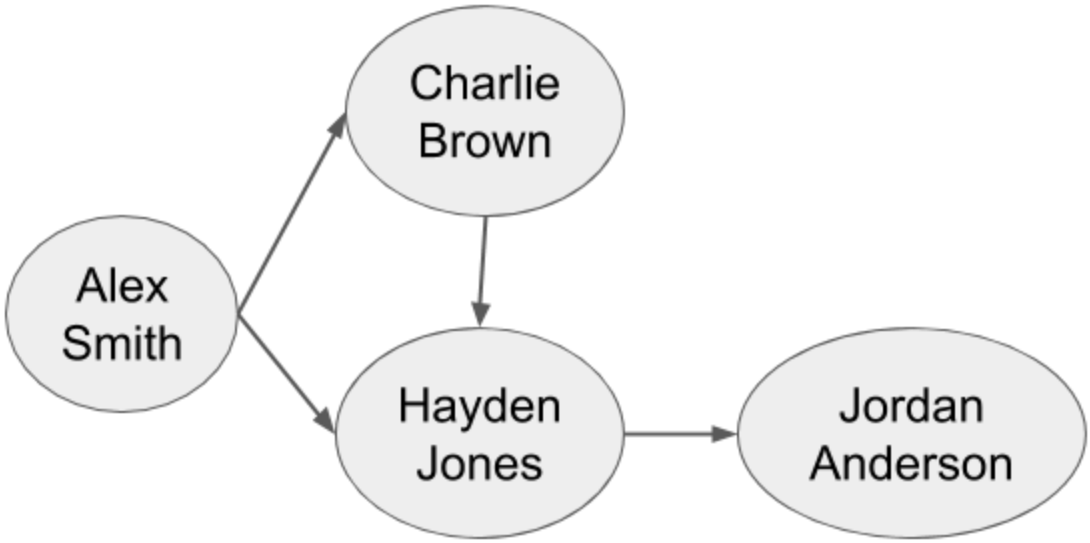
more info: https://www.futurelearn.com/courses/data-management-and-visualisation-data-fundamentals/3/steps/1600118

using SPARQL on DBPedia

- pronounced 'sparkle', its like SQL but for RDF data
- DBPedia a community-created database version of Wikipedia
- example for Tom Hanks: https://github.com/suzannelittle/ca682i/blob/master/resources/1_3_9_linked_data.md

Non-tabular formats: graph

- graph databases (from a Neo4j developer)

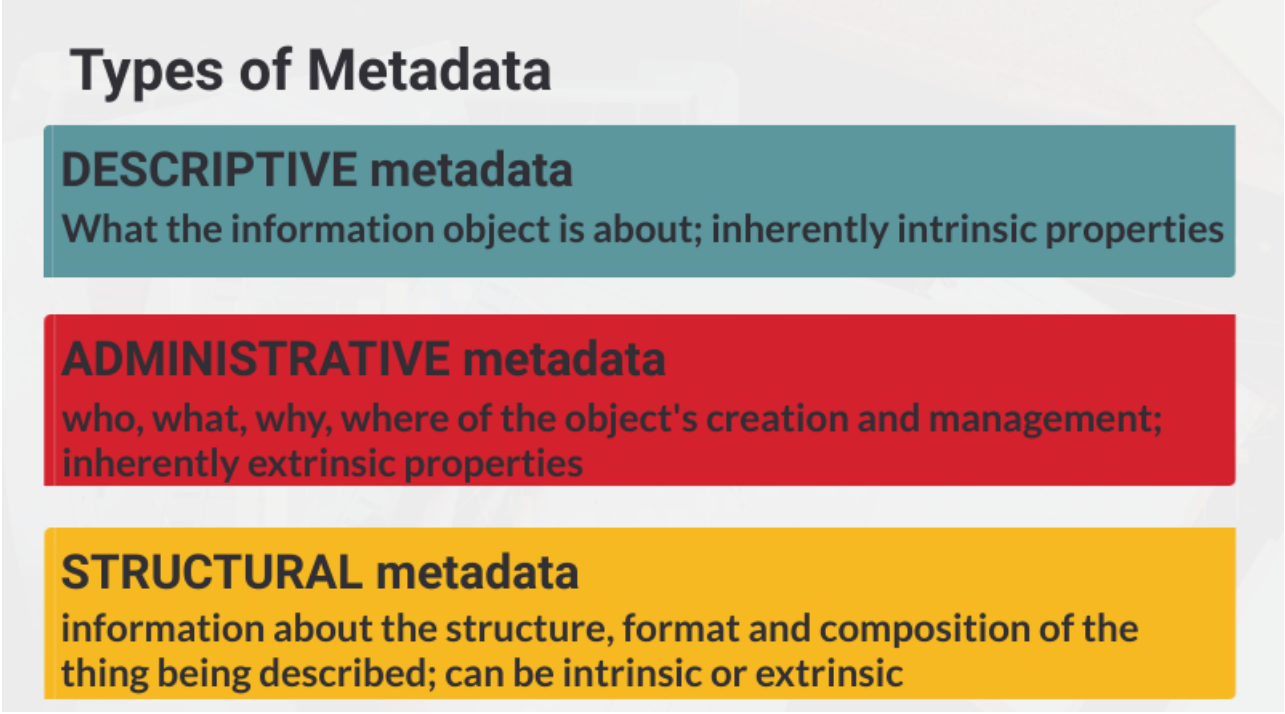


---

wk4

Metadata

- metadata is also the place where information on rights management and privacy are often included.



- EXIF is metadata for photos
  - descriptive meta: 'leaning tower of pizza' or 'my holiday' label (many meanings based pon context)
  - EXIF standard from cameras (DateTimeOriginanl, ApertureValue)

Metadata can apply to different dimensions or concepts:

- Abstraction: how **close** the metadata is to the data.
- Granularity: how **detailed** the metadata is.

Abstraction (how close to the data) is one way that metadata can be structured. When dealing with complex objects, especially digital representations of physical objects, you can see where there could be different ways of describing the data. Another way of thinking of this is the **granularity** (how detailed) of the metadata descriptions.

https://www.futurelearn.com/courses/data-management-and-visualisation-data-fundamentals/3/steps/1600128

How is metadata created:

> 1. **Simple:** This is metadata that is often generated automatically during the creation of a digital object. Recall in the introduction to this topic, we looked at the embedded metadata in digital photographs. Eg, EXIF, document headers, time stamps.
> 2. **Structured:** This is metadata that adheres to a standard. It may also include manually entered data or conventions. Example: Dublin Core.
> 3. **Professional:** Librarians, Archivists, Curators are professional metadata creators! They apply standards and are trained to create consistent, independent and high-quality metadata. Eg, the Library of Congress standards.
> 4. **Crowd Sourced:** Social media content will have examples of simple and structured metadata associated with it but also crowdsourced labels. These may gradually evolve into accepted standard labels from comments or hashtags. Example: hashtags, comments.

So how much metadata do we need?

Alemu & Stevens (2015) describe **four principles** that may help in deciding how much metadata is required.

- The principle of sufficiency and necessity;
- The principle of user convenience;
- The principle of representation;
- The principle of standardisation.

Fundamentally, it's important to consider the tradeoffs between:

- organisation (adding, duplicate detection, storage) and
- retrieval (query, search).

problems:

Read this 2001 article by author **Cory Doctorow** "Metacrap: Putting the torch to seven straw-men of the meta-utopia".

In 2001 there was a particular push for exhaustive, high-quality, machine-readable metadata in the belief that it would create enormous opportunities for new services. In the article, Doctorow looks at seven obstacles that prevent this "meta-utopia":

- People lie.
- People are lazy.
- People are stupid.
- People delude themselves.
- Schemas aren't neutral.
- Metrics distort or limit.
- There's more than one way!

Some of these ideas will come up again in the second course when we look at data cleaning, as metadata created by humans can exhibit a lot of inconsistencies and errors.

A contrasting view on metadata is presented in the Introduction to "An emergent theory of digital library metadata: enrich then filter" by **Alemu & Stevens (2015)**.

Metadata is not only about a single object, either physical or digital, but can also be applied to a datase

w4 quiz: https://www.futurelearn.com/courses/data-management-and-visualisation-data-fundamentals/3/steps/1600123/quiz/introduction

In [ ]: