TECHNOLOGY
NETWORKS

DATA AND DATA INTEGRITY

In the last few decades, there has been an increase in the application of big data across many fields, and significant changes have been made as to how data is acquired, analyzed and visualized.

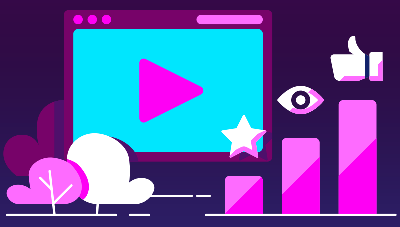
In this infographic, we define big data and what is meant by data integrity.

WHAT IS BIG DATA?

One definition defines big data by the 3 "V"S

VARIETY

In Big Data, a mix of structured and unstructured data is used – like the difference between a video's click number, and the content of that video.



VELOCITY

Data comes in extremely fast, often in real time.

VOLUME

There is an abundance of data.

BUT SOMETIMES OTHER FACTORS ARE ADDED

INCOMPLETENESS

INCONGRUENCY

COMPLEX REPRESENTATION

SOURCE HETEROGENEITY

MULTISCALE DATA FORMAT

All these factors mean that Big Data can be pretty tough to visually represent.

HOW DO WE VISUALIZE BIG DATA?

To work out how to visualize your data, you must first work out what type of data you have. Let's consider some types using an (refreshingly non-technical) example: the number of giraffes in a zoo.

TYPES OF DATA

UNIVARIATE

Where there is just one variable present – such as giraffe number over time.

TWO-DIMENSIONAL DATA

Where two variables are considered, such as giraffe number vs. number of zookeepers.

MULTIDIMENSIONAL DATA

Generally used to mean data where three or more variables are involved, so giraffe number vs. total number of giraffe spots vs. height of giraffes.

TEXTS AND HYPERTEXTS

A webpage listing the names of all the giraffes.

HIERARCHICAL DATA

A graph showing structure and relationship of data – in this instance, it could be a giraffe genealogy map.

WHAT IS DATA INTEGRITY?

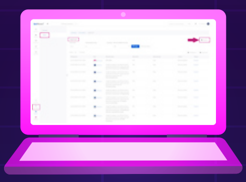
Data integrity has become an important target for companies across biotech, pharma and biomedical research.

The FDA defines data integrity as a measurement of the completeness, consistency and accuracy of data. The data should meet the criteria of ALCOA principles.

WHAT CAUSES A REDUCTION IN DATA INTEGRITY?

1

DATA RECORDS DON'T COMPLY TO STANDARDS



Inspectors assessing the integrity of data will want to see that that it meets the ALCOA principles. Data that is located in messy paper notebooks will not meet these standards. Neither will poorly-organized excel sheets. Instead, electronic data should be linked together and easy-to-follow. An audit trail can help you get your data in shape.

2

A LACK OF BACKUPS

A backup is defined by the FDA as a "true copy of the original record that is maintained securely throughout the record retention period". Backing up data to multiple sources will guarantee the safety of your backups and the integrity of your data.

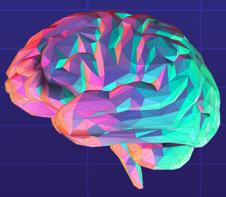
3

SHARED LOGINS

The FDA have identified shared logins as a common way that parts 211 and 212 of the CGMP requirements are breached. Document controls must ensure that all actions can be attributable to a specific individual.

4

LACK OF TECHNICAL KNOWLEDGE



A recent survey of data integrity expert conducted for Technology Networks identified that cultural reasons within companies may be a significant contributor towards data integrity issues. If company leaders perceive data integrity to be just an "IT problem", it is unlikely individuals will be given the proper support and training to tackle integrity issues.

Reference:

https://journalofbigdata.springeropen.com/articles/10.1186/s40537-015-0031-2