

# Introduction to COMP3008: Big Data Analytics

## COMP3008

Lauren Ansell

# Introduction To The Module

COMP3008 introduces current and emerging practices for dealing with large-scale database systems.

In this module, you will work with both structured and semi-structured datasets and choose appropriate storage structures and analytical methods for them.

A representative of recent nonrelational trends will be presented - namely, graph databases.

# Module Content: Graphical Databases

Graph databases: A collection of nodes and edges to record data and relationships.

Graphical databases highlight the relationships between data.

Recommender systems: Computer systems that make recommendations.

# Module Content: Handling Big Data

Big data: Work with datasets that are too large or complex to handle with traditional data-processing applications.

Topics in big data:

- document classification
- topic mining
- visualisation
- prediction

# Assessment

Coursework - no exams or tests.

Coursework released and returned through the DLE.

Formative feedback available before the submissions.

If something (for example, a health issue) causes a problem with assessment apply for extenuating circumstances - ECs

# Coursework Details

The coursework consists of two parts:

- Set Exercises - 30%
- Report - 70%

Details for each part of the coursework can be found on the DLE under "Assessments".

Both parts are to be submitted through the DLE.

Deadlines:

- Set Exercise - 3rd March 2026 3pm
- Report - 5th May 2026 3pm

Robinson, I., Webber, J. and Eifrem, E., 2015. *Graph Databases: New Opportunities for Connected Data*. O'Reilly Media, Inc.

Needham, M. and Hodler, A.E., 2021. *Graph Data Science For Dummies*. John Wiley & Sons, Inc.

Baumer, B., Kaplan, D. and Horton, N., 2021. *Modern Data Science with R*. CRC Press.

Hvitfeldt, E. and Silge, J., 2022. *Supervised Machine Learning for Text Analysis in R*. CRC Press.

## Linkedin Course: Database Clinic: Neo4j

[https://www.linkedin.com/learning-login/share?forceAccount=false&redirect=https%3A%2F%2Fwww.linkedin.com%2Flearning%2Fdatabase-clinic-neo4j%3Ftrk%3Dshare\\_ent\\_url](https://www.linkedin.com/learning-login/share?forceAccount=false&redirect=https%3A%2F%2Fwww.linkedin.com%2Flearning%2Fdatabase-clinic-neo4j%3Ftrk%3Dshare_ent_url)

## Neo4j reference card

<https://neo4j.com/docs/cypher-cheat-sheet/5/auradb-enterprise/>

## GitHub repository:

<https://github.com/laurenansell/COMP3008>

# Introduction To Today's Session

Today's topics:

- What is big data?
- What are the properties of big data?
- Importing large datasets into Neo4j.
- How to update and index data in a Neo4j database.

Session learning outcomes - by the end of today's lecture you will be able to:

- Explain the what we mean by big data.
- Explain the 5 V's of big data.
- Know how to create records in Neo4j.

# Big Data

Big data refers to datasets whose size is beyond the ability of typical database software tools to capture, store, manage, and analyse. - John Morton (Chief Technology Officer, SAS UK).

This definition does not specify how large (terabytes, petabytes, zettabytes) a dataset has to be in order to be considered big data. It assumes that, as technology advances, the size of datasets that qualify as big data will also increase.

Also the definition may vary by sector, depending on what kinds of software tools are commonly available and what sizes of datasets are common in a particular industry.

- Big data primarily refers to data sets that are too large or complex to be dealt with by traditional data-processing application software.
- Data with many entries offer greater statistical power, while data with higher complexity may lead to a higher false discovery rate.
- Though used sometimes loosely partly due to a lack of formal definition, the best interpretation is that it is a large body of information that cannot be comprehended when used in small amounts only.
- Big data philosophy encompasses unstructured, semi-structured and structured data; however, the main focus is on unstructured data.

# An Experiment....

In January 2017, we retrieved a dataset composed of New Year resolutions published on Twitter. The dataset was collected over 10 consecutive days and consisted of 52,583 tweets.

On the same 10 days when we collected 52,583 New Year Resolutions published on Twitter in January 2017, we collected 13,025,592 tweets about Donald Trump.

# The History of Big Data

The term “Big Data” was popularised in the 1990s by John Mashey.

The “size” where we begin classifying data as Big Data has been constantly changing, as of 2012 ranging from a few dozen terabytes to many zettabytes of data.

In 2018, Big Data was defined as “where parallel computing tools are required to handle the data”.

What this definition represents is a distinct and clearly defined change in the computer science used.

# Challenges in Big Data

Challenges in Big Data analysis include:

- Capturing the data
- Storing the data
- Analysing the data
- Visualising the data
- Predictions from the data

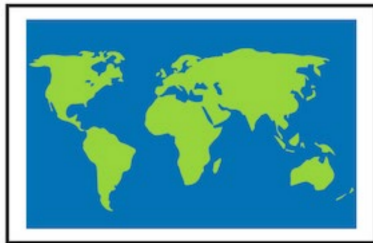
# Data v Information

## DATA

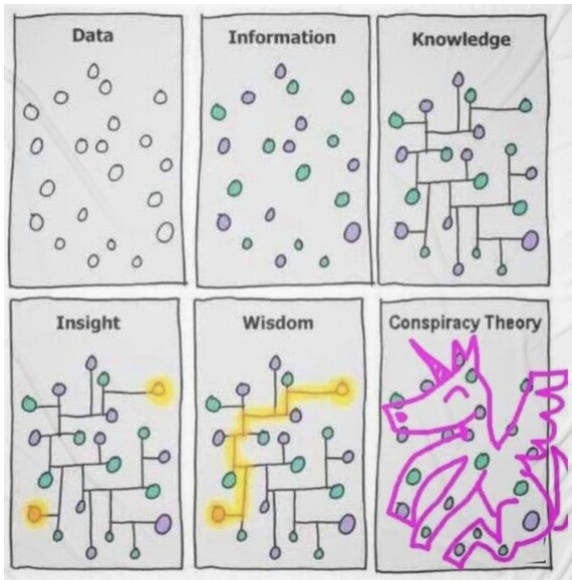


Data is raw, unorganised facts that need to be processed.

## INFORMATION



Information refers to the meaning of data as understood by some user.



# Big Data: The V's

Big data is often described using the five Vs:

- 1 Volume
- 2 Velocity
- 3 Variety
- 4 Veracity
- 5 Value

According to the Gartner IT glossary, big data is “high-volume, high-velocity and high-variety information assets that demand cost-effective, innovative forms of information processing for enhanced insight and decision making”.

Volume refers to the vast amounts of data continuously generated.

We are not talking Terabytes but Zettabytes.

We create 2.5 quintillion bytes of data everyday (enough to fill 10 million Blu-ray discs).

Velocity refers to the speed at which new data is generated and the speed at which data moves around - for example, social media messages going viral in seconds.

Every 60 seconds there are: 72 hours of footage uploaded to YouTube; 216,000 Instagram posts; 204,000,000 emails sent.

Technology allows us now to analyse the data while it is being generated (sometimes referred to as in-memory analytics), without ever putting it into databases.

Variety refers to the different types of data that we can now use.

In the past we only focused on structured data that neatly fitted into tables of relational databases, such as financial data.

Today, 80% of the world's data is unstructured.

With big data technology we can now analyse and bring together data of different types such as messages, social media conversations, photos, sensor data, video or voice recordings.

# Veracity

Veracity refers to the messiness or trustworthiness of the data.

With many forms of big data, quality and accuracy are less controllable.

For example, in X posts:

- abbreviations
- typos
- colloquial speech
- reliability of content
- accuracy of content

Technology now allows us to work with this type of data.

It is estimated that poor data quality costs the US economy 3.1 trillion US dollars per year (IBM Big Data & Analytics Hub).

Value refers to the ability to achieve greater value through insights from the analysis of big data.

Aircraft engine manufacturers make use of big data analysis to predict engine events that lead to costly airline disruptions with 97% accuracy.

If this prediction capability had been available in the previous year, it would have saved \$63 million.

## **Exhaustive**

Whether the entire system is captured or recorded or not. Big data may or may not include all the available data from sources.

## **Fine-grained and uniquely lexical**

Respectively, the proportion of specific data of each element per element collected and if the element and its characteristics are properly indexed or identified.

## **Relational**

If the data collected contains common fields that would enable a conjoining, or meta-analysis, of different data sets.

## **Extensional**

If new fields in each element of the data collected can be added or changed easily.

## **Scalability**

If the size of the big data storage system can expand rapidly.

# What is Neo4j?

Neo4j is a graph database management system developed by Neo4j, Inc.

The data elements Neo4j stores are nodes, edges connecting them, and attributes of nodes and edges.

Nodes and edges can be labelled and labels can be used to narrow searches.

# Cypher Query Language (CQL)

Cypher is Neo4j's query language.

Cypher is a declarative, SQL - inspired language for describing patterns in graphs.

# CREATE command syntax

```
CREATE (<variable-name>:<label-name>)
```

Syntax Element	Description
CREATE	Neo4j CQL command
<node-name>	name of the node that we are going to create
<label-name>	name of a label

Commit means updating a record in a database.

In the context of a database transaction, commit refers to the saving of data permanently after a set of changes.

A commit ends a transaction within a database and allows all users to see the changes.

## Big data

- Definition of big data has changed over time.
- Big data is described by the 5 V's.

## Neo4j

- Neo4j is a graph database management system.
- Neo4j uses Cypher Query Language.