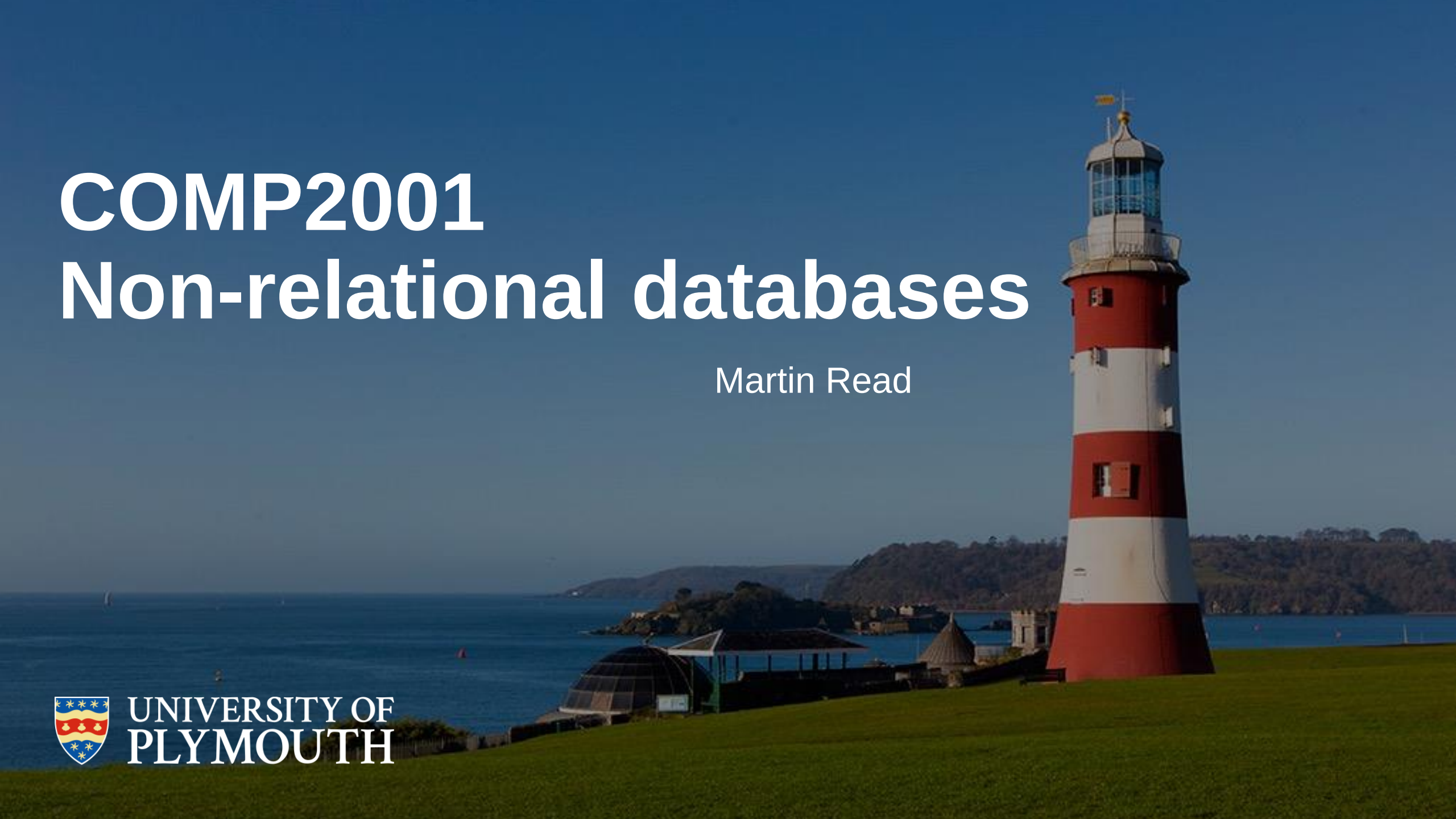


COMP2001
Non-relational databases

Martin Read

UNIVERSITY OF PLYMOUTH

Session outline

- To introduce landscape for advanced database concepts and information retrieval
- Summarise major approaches of the NoSQL data model and how they differ from the relational model

UNIVERSITY OF PLYMOUTH

Databases

“A database is a collection of information that exists over a long period of time, often many years. ”

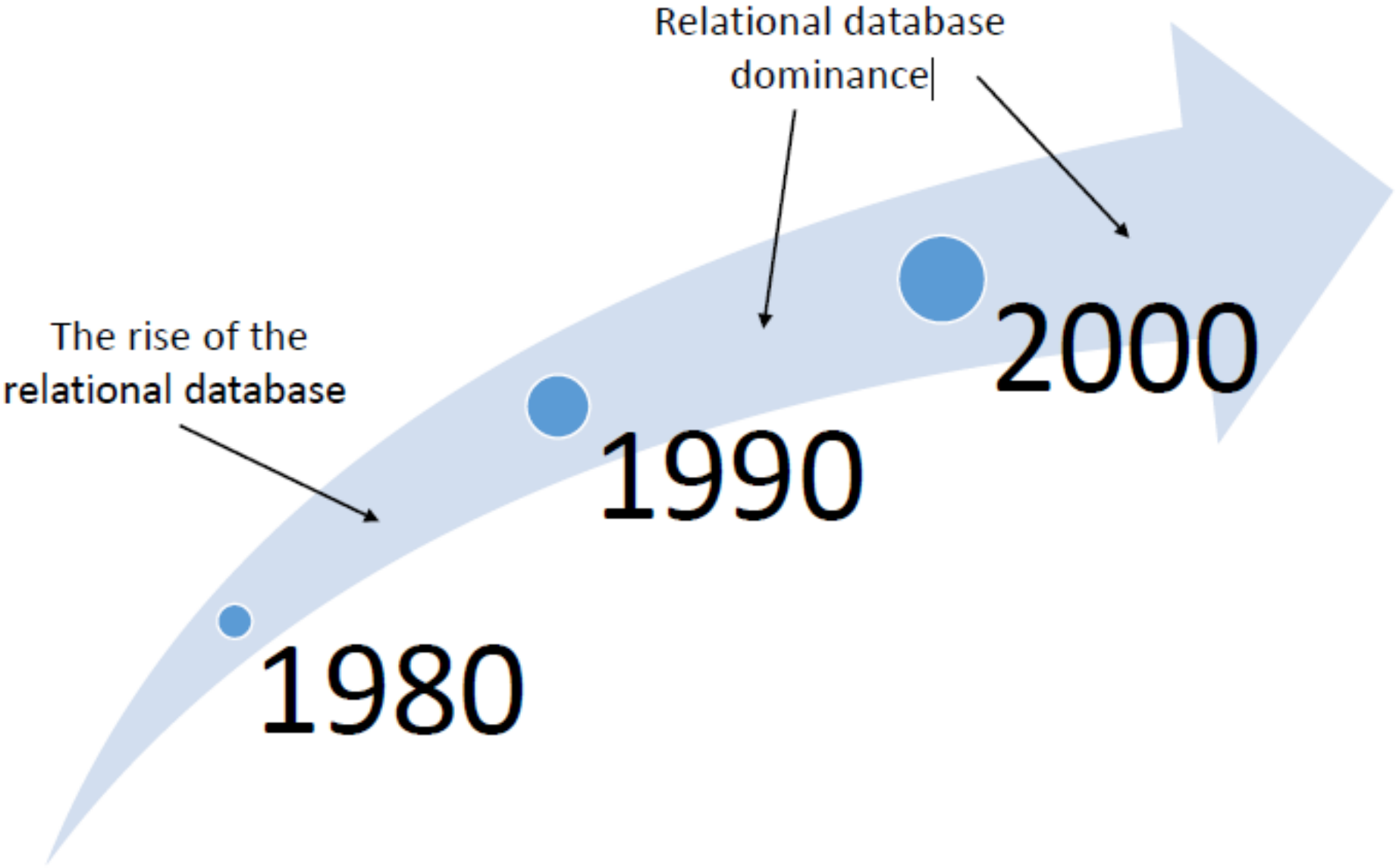
– *Garcia-Molina et al. Database Systems: The Complete Book*

Databases are powerful tools for doing exactly what computers do best:

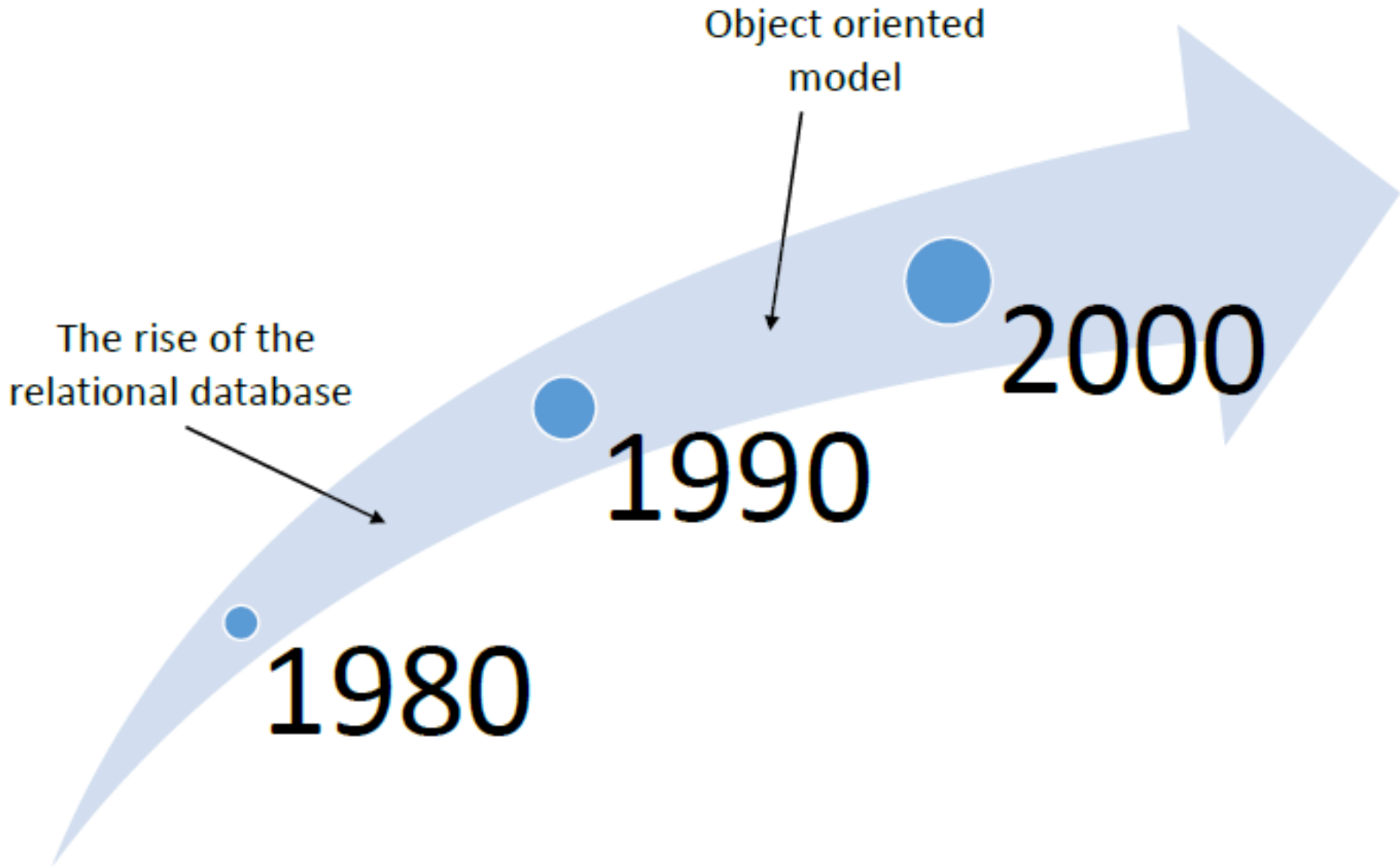
- store
- manipulate
- display information

UNIVERSITY OF PLYMOUTH

Background (history)



Background (history)



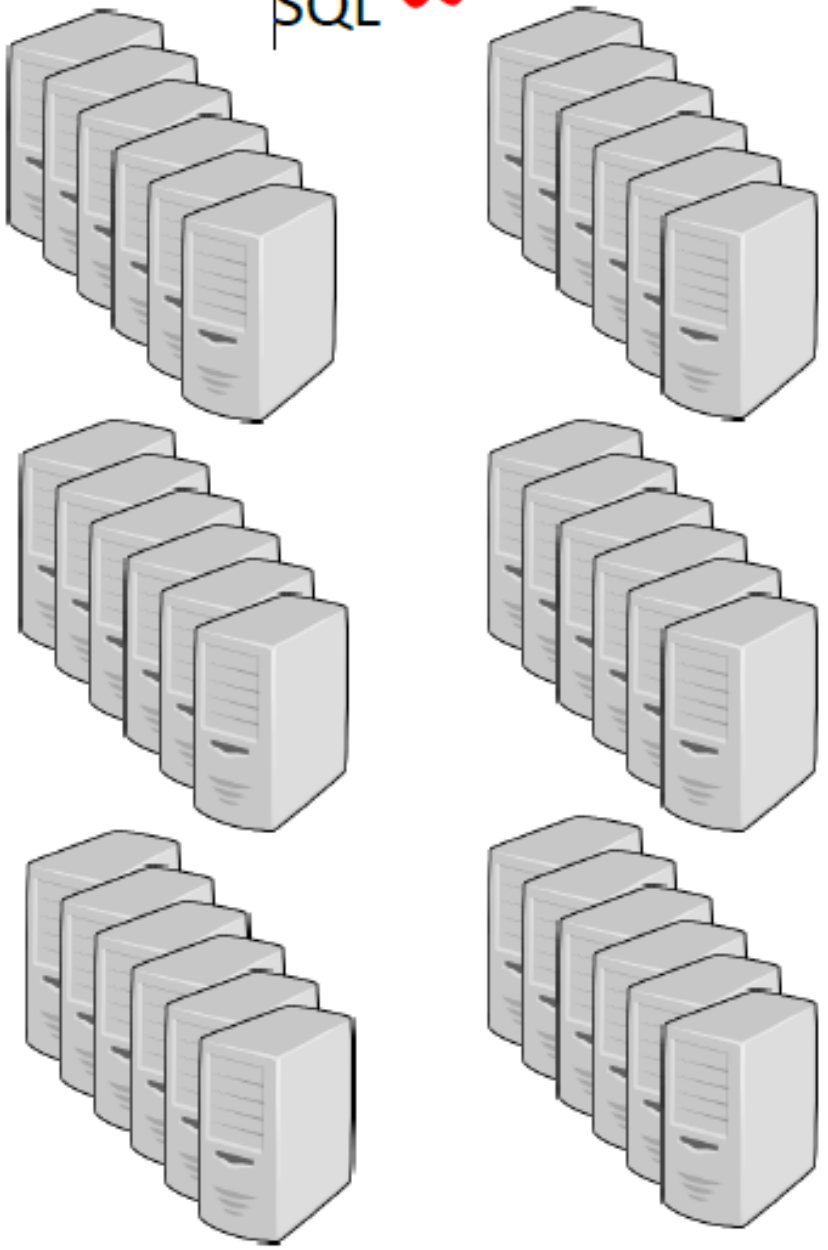
Internet traffic

Servers

SQL ✓

Unnatural acts...

SQL ✗



New data storage systems

UNIVERSITY OF PLYMOUTH

Origins

Google	Apache Software Foundation
Google File System (GFS)	Hadoop Distributed File System (HDFS)
Google MapReduce	MapReduce
Google BigTable	Apache HBase

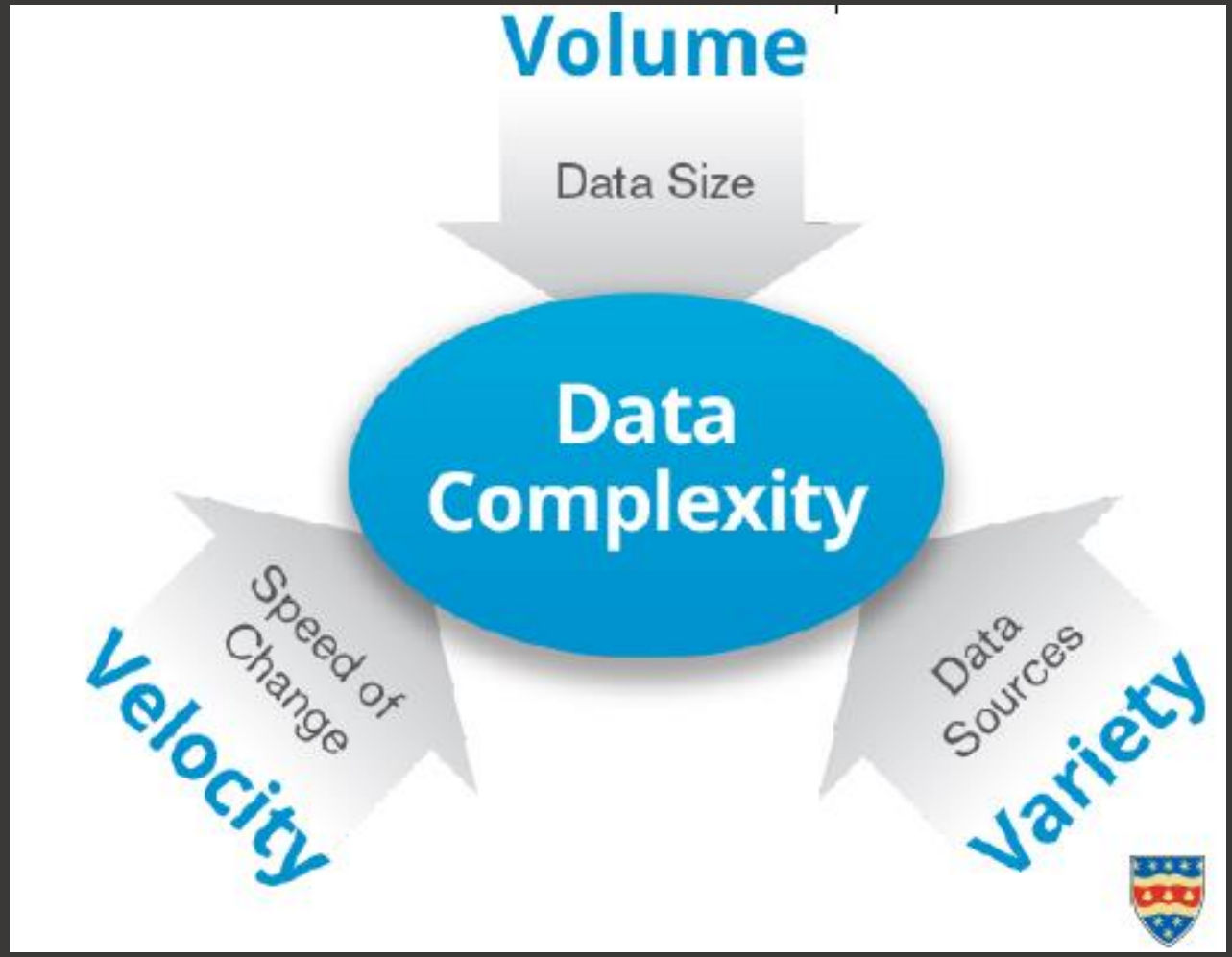
Google whitepapers

The origins of non-relational databases can be traced back to 3 Google whitepapers:

- Sanjay Ghemawat, Howard Gobioff, and Shun-TakLeung. 2003. **The Google file system**. In Proceedings of the 19th ACM symposium on Operating systems principles(SOSP '03). ACM, New York, NY, 29-43.
- Jeffrey Dean and Sanjay Ghemawat. 2010. **MapReduce: A flexible data processing tool**.Commun. ACM53, 72-77.
- Fay Chang, *et al*. 2008. **Bigtable: A Distributed Storage System for Structured Data**. ACM Trans. Comput. Syst.26, 2

Big Data

- Big data is often described using the three Vs: **volume**, **velocity** and **variety**



UNIVERSITY OF PLYMOUTH

Four Vs

Sometimes big data is described using four Vs:

- Volume
- Velocity
- Variety
- **Veracity**

UNIVERSITY OF PLYMOUTH

Veracity

Refers to the messiness or trustworthiness of the data

- **With many forms of big data, quality and accuracy are less controllable**
- Estimated that poor data quality costs the US economy 3.1 trillion US dollars per year
  - consider Twitter posts, abbreviations, typos & colloquial speech
  - reliability and accuracy of content
- technology now allows us to work with this type of data

Five Vs

Sometimes big data is described using five Vs:

- Volume
- Velocity
- Variety
- Veracity
- **Value**

UNIVERSITY OF PLYMOUTH

Value

- 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. This could save millions

Criticism

- **Big data** is an ill-defined term
- There is a lack of precision in the definition of big data, which leads to ambiguity

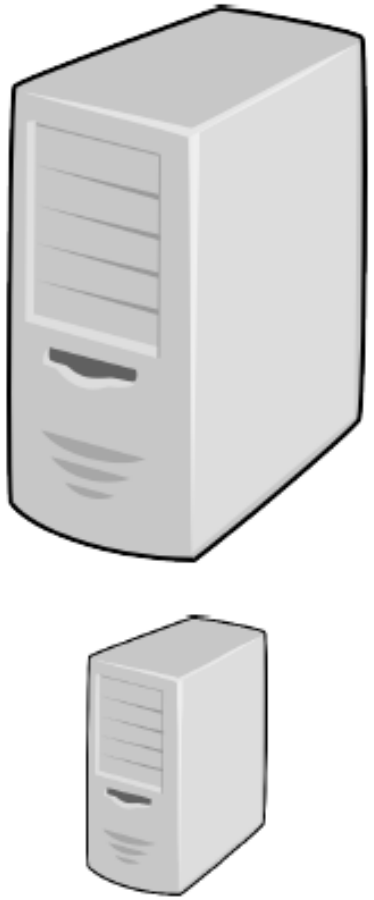
- **NoSQL** is an accidental neologism
- a term with no prescriptive definition. All you can make is an observation of common characteristics

UNIVERSITY OF PLYMOUTH

New data storage systems

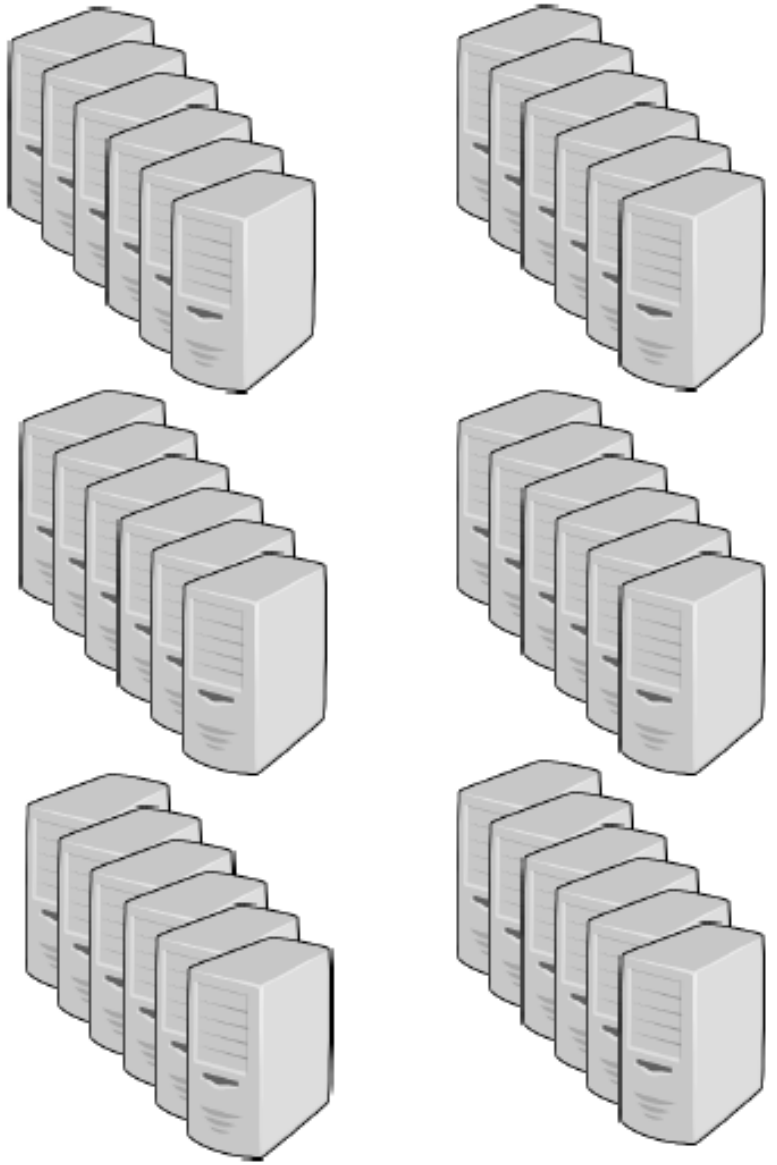
UNIVERSITY OF PLYMOUTH

Scaling



Scale up

Scale out



UNIVERSITY OF PLYMOUTH

Emerging data models

- **Bigtable:** Google Bigtable
- **Key-value store**, or **key-value database:** Redis, MemcacheDB, Berkeley DB (BDB), HamsterDB...
- **Document-oriented database**, or **document store**: MongoDB, CouchDB, OrientDB, RavenDB, Lotus Notes....
- **Graph database**: Neo4j, HyperGraphDB...

NoSQL... ?

- NoSQL was a Twitter hashtag (#nosql) chosen for a meetup organised by **Johan Oskarsson**in San Francisco in 2009 to discuss new databases

- “NoSQL is an accidental term with no precise definition”. [Sadalage& Fowler: NoSQL Distilled, 2012]

NoSQL

- **NoSQL**means **Not Only SQL**
- implying that when designing a software solution or product, there is more than one storage mechanism that could be used

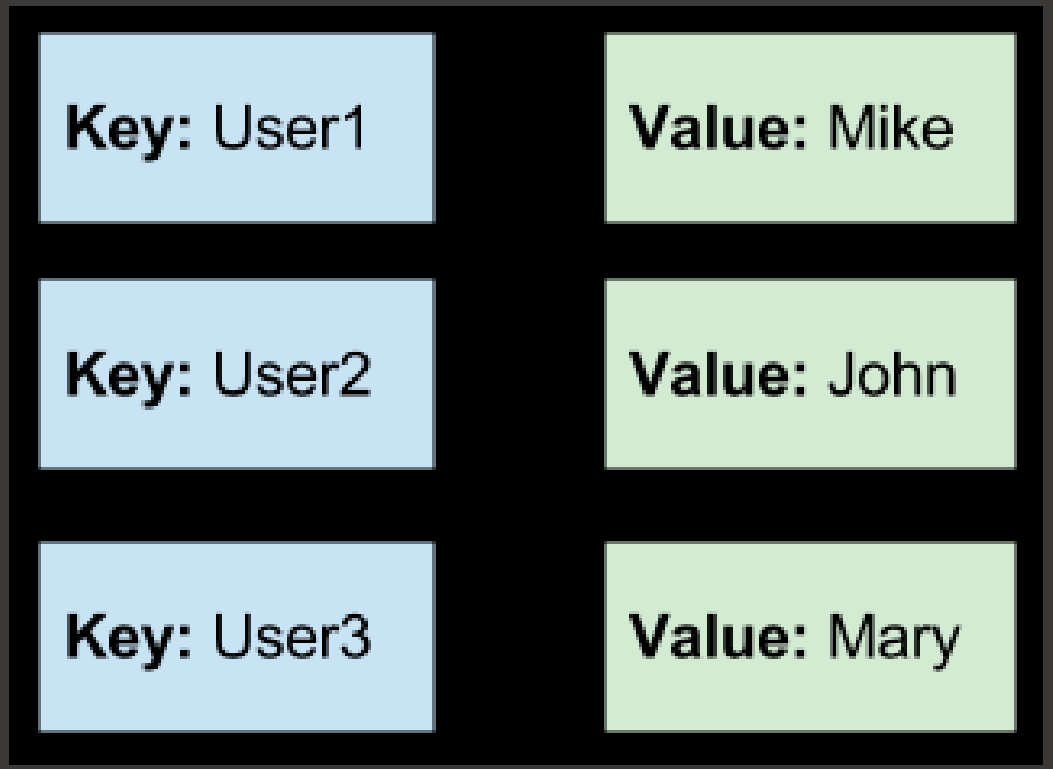
NoSQL features

- Non-relational
- Cluster friendly
- Open source
- 21st Century Web

UNIVERSITY OF PLYMOUTH

Key-value store

- Key-value pairs are similar to a table
  - key serves as an index to find an associated value
- Key-value pairs are similar to accessing data in memory
  - the key is a **memory location** & the value is the **data stored at the location**, making key-value pairs a good data model for **in-memory databases**




Any Questions?
UNIVERSITY OF
PLYMOUTH