Advanced databases

Graham Kemp | Data Science and AI Division | 2023-02-23

Guest lecture introducing the course: “Advanced databases” (DAT475 | DIT930)

Which area is most glamorous?

A. High-performance computing

B. Computer graphics

C. Machine learning

D. Databases

Financial Times Survey: November 27 2001

“Data integration and management is an area with less glamour than high-performance computing ...”

“... but, probably, more practical relevance for the biotech industry. Researchers need to organise and integrate information about genes and proteins from many different sources, in many formats and file types, so that they can uncover patterns and associations."

Financial Times Survey: November 27 2001

“Data integration and management is an area with less glamour than high-performance computing ...”

“... but, probably, more practical relevance for the biotech industry. Researchers need to organise and integrate information about genes and proteins from many different sources, in many formats and file types, so that they can uncover patterns and associations."

“We Don't Need Data Scientists, We Need Data Engineers”, January 2021

- “There are **70% more open roles** at companies in *data engineering* as compared to *data science*. As we train the next generation of data and machine learning practitioners, let’s place more emphasis on engineering skills.”

- *“Data scientist*: Use various techniques in statistics and machine learning to process and analyse data. Often responsible for building models to probe what can be learned from some data source, though often at a prototype rather than production level.”
- *“Data engineer*: Develops a robust and scalable set of data processing tools/platforms. Must be comfortable with SQL/NoSQL database wrangling and building/maintaining ETL pipelines.”

Senior Machine Learning Scientist @ Amazon Alexa AI

https://www.mihaileric.com/posts/we-need-data-engineers-not-data-scientists/

“Data science is different now”, February 2019

- “What do you think of when you read the phrase ‘data science’? It’s probably some combination of keywords like statistics, machine learning, deep learning, and ‘sexiest job of the 21st century’. Or maybe it’s an image of a data scientist, sitting at her computer, putting together stunning visuals from well-run A/B tests. Either way, it’s glamorous, smart, and sophisticated.”

Machine learning engineer at Tumblr

https://vickiboykis.com/2019/02/13/data-science-is-different-now/

“Data science is different now”, February 2019

- “What do you think of when you read the phrase ‘data science’? It’s probably some combination of keywords like statistics, machine learning, deep learning, and ‘sexiest job of the 21st century’. Or maybe it’s an image of a data scientist, sitting at her computer, putting together stunning visuals from well-run A/B tests. Either way, it’s glamorous, smart, and sophisticated.”

- “I recommend learning SQL for everyone, regardless of whether their ambition is to be a data engineer, ML expert, or AI superwhiz.”
- “SQL is not sexy, ... But for all intents and purposes, in order to understand how to access data, chances are extremely high that you’ll come across a database somewhere that will require you to write some SQL queries and get an answer.”

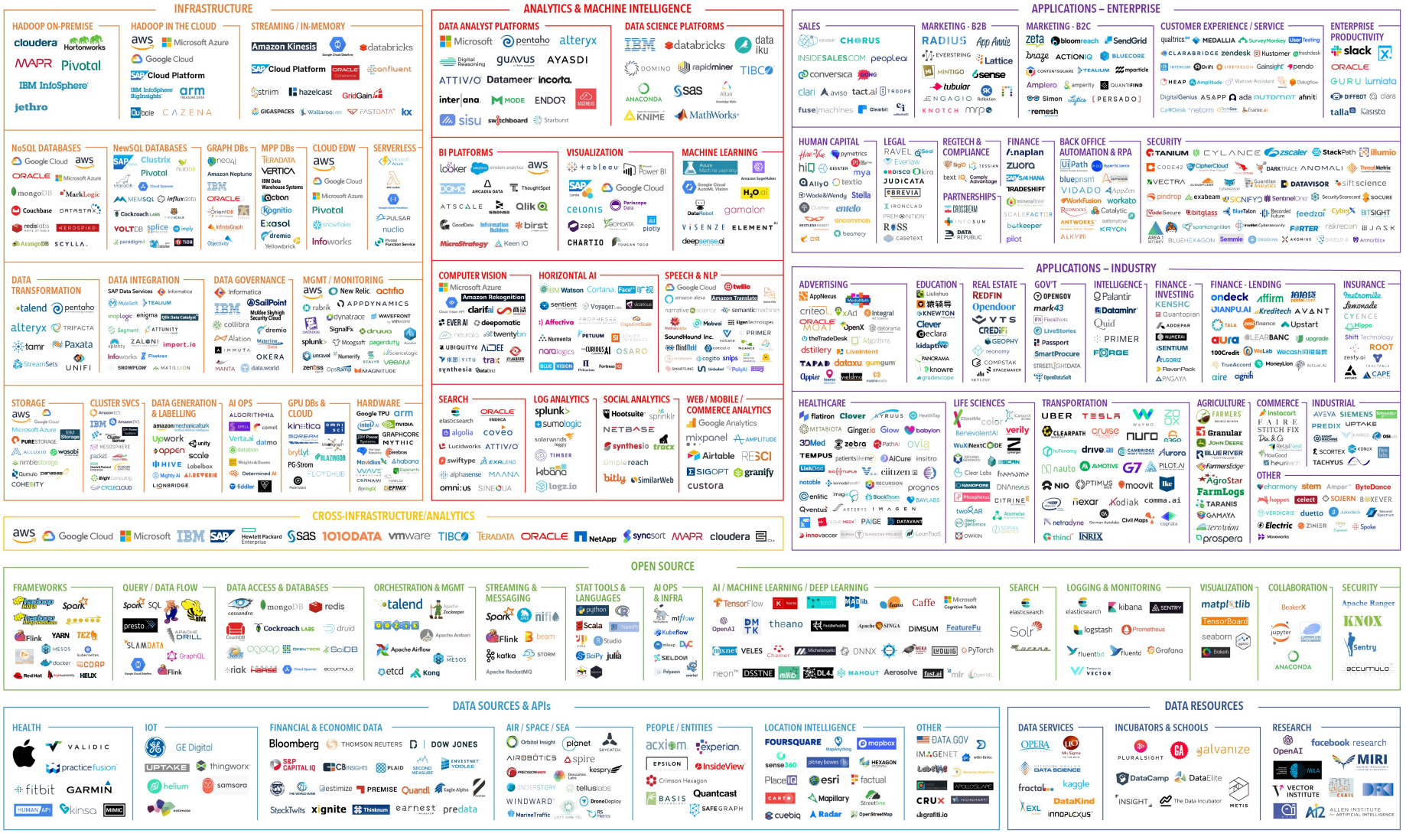
https://vickiboykis.com/2019/02/13/data-science-is-different-now/

Advanced databases (DAT475|DIT930)

- Given for the first time in 2022
- Study period 4
- 7.5 credits

An aside …

DATA & AI LANDSCAPE 2019

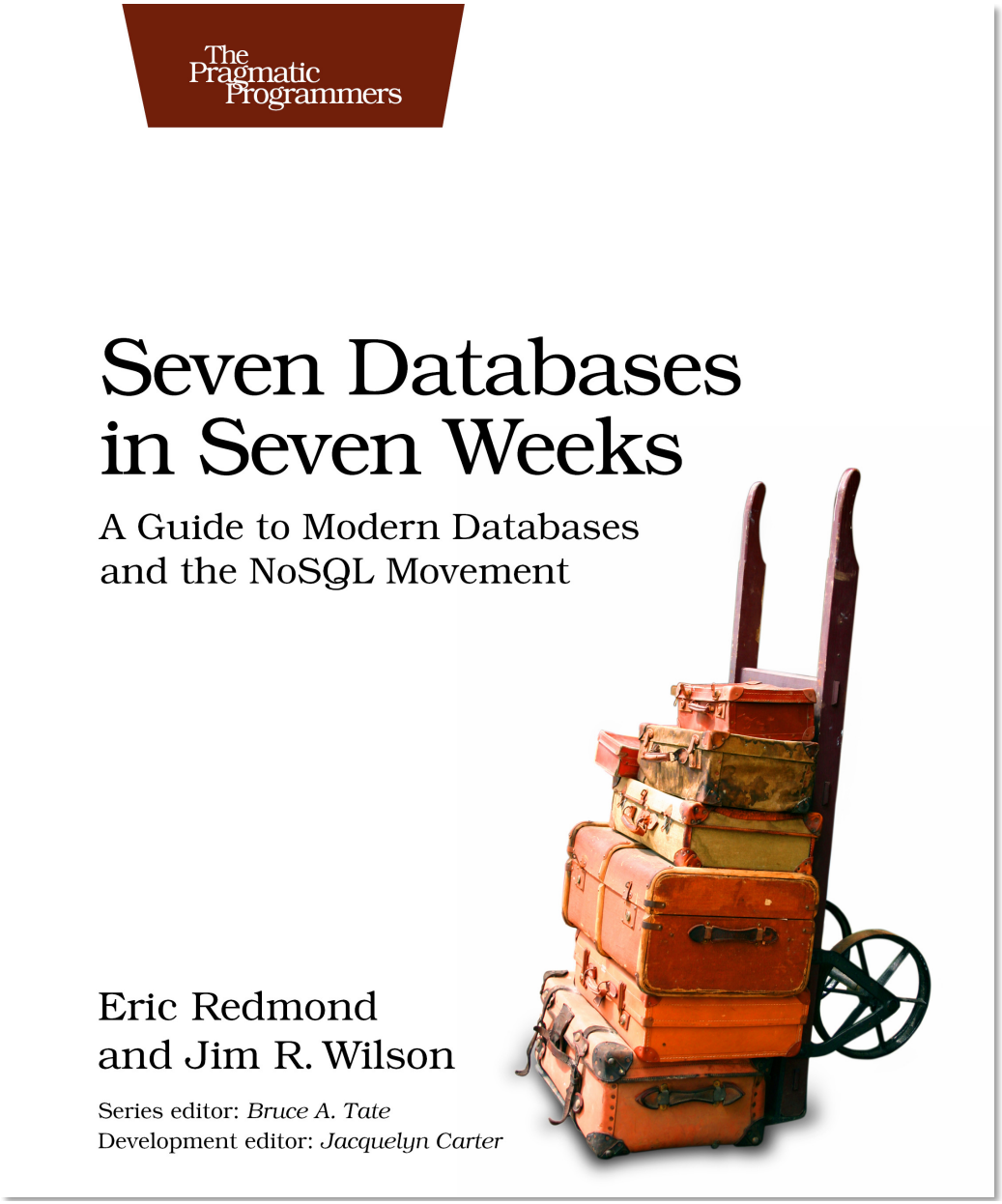


FIRSTMARK
EARLY STAGE VENTURE CAPITAL

Data and AI landscape

- For several years Matt Turck has published a graphical overview of the Big Data and AI landscape, and the lecture slides include a recent edition. Only some technologies and companies are featured in the (2019) picture, selected form a list of 1335. A more recent (2020) picture is based on a list of 1479 technologies.
- It would be impossible to try to introduce so many technologies in a course.
- Even gaining meaningful hands-on experience with seven different databases in these few weeks would be difficult.

- http://dfkoz.com/ai-data-landscape/



A book about NoSQL systems that features 1 relational DBMS and 6 NoSQL systems:

- PostgreSQL
- Riak
- HBase
- MongoDB
- CouchDB
- Neo4J
- Redis

A good book for learning about some NoSQL systems, but the "Advanced databases" course is not only about NoSQL.

Advanced databases (DAT475|DIT930)

Course contents:

- database management system architecture and implementation
- concurrency and recovery
- indexes
- query processing and optimization
- Semantic Web; RDF; RDF Schema; SPARQL
- ontologies
- NoSQL systems; aggregation-orientation; CAP theorem
- querying graph databases
- database applications

Learning outcomes

Knowledge and understanding

- describe concepts relating to the implementation of database management systems
- compare and contrast features of relational and non-relational database management systems

Skills and abilities

- construct Web ontology language statements corresponding to an Entity-Relationship diagram
- construct RDF (Resource Description Framework) triples that contain data for a given domain
- implement a graph database for a given domain
- retrieve data using declarative query languages for graph databases

Judgement ability and approach

- discuss advantages and disadvantages of different database design decisions
- discuss advantages and disadvantages of alternative query plans
- discuss suitability of different database management systems for various tasks

Not only new stuff

- Database management systems have been around for over 50 years.
- For most of that time relational database systems have been dominant, but different kinds of non-relational systems have been in and out of fashion throughout that time, and today many applications are better supported by a variety of non-relational systems that have appeared in recent years.
- While we will look at some recent and emerging technologies in this course, we shall also look back at some influential ideas and trends since, to understand today's DBMS landscape, and to be prepared for future developments, it's useful to understand how we got here.

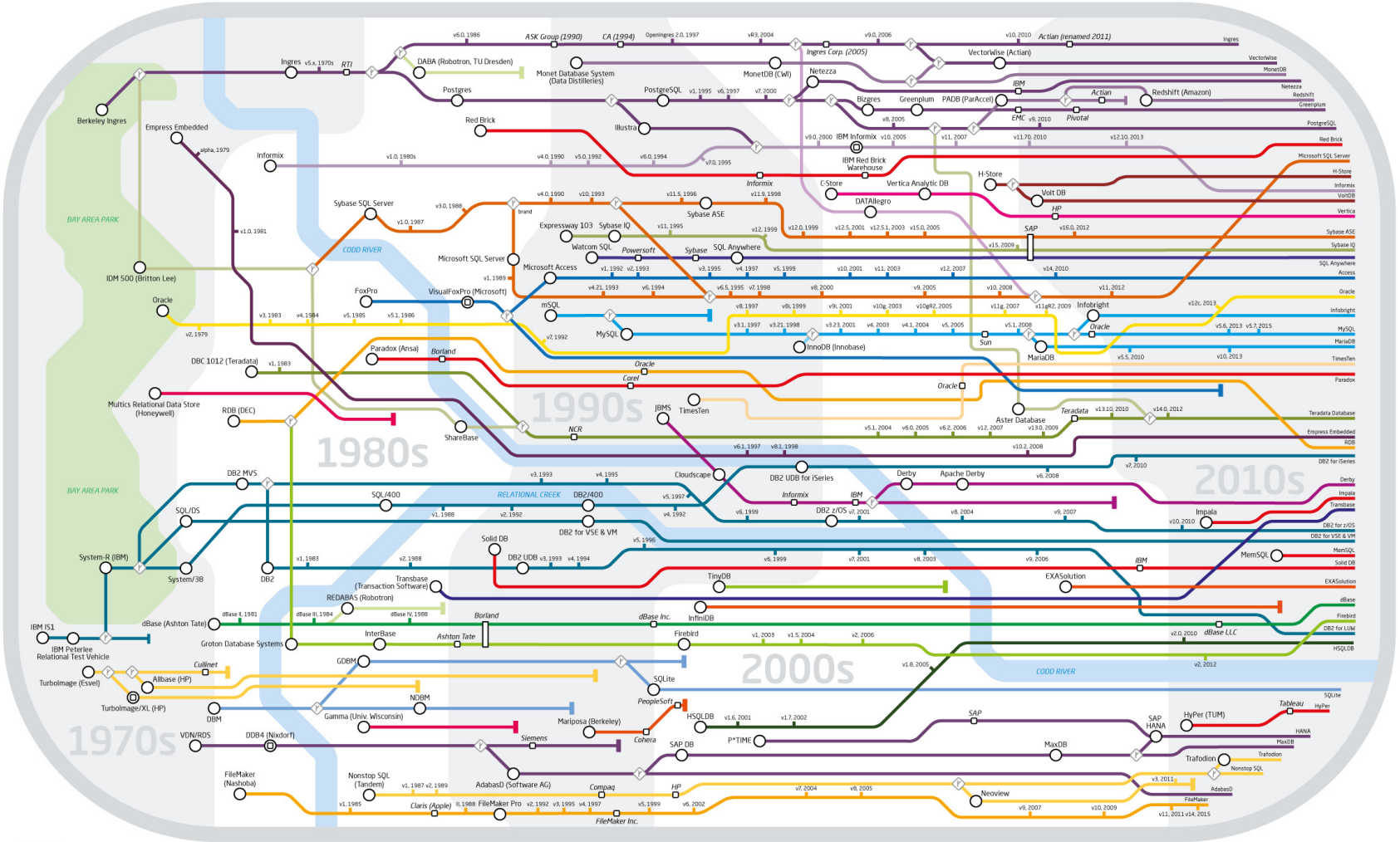
The prerequisites for this course include at least 7.5 credits in databases – why?

- It is assumed that you are familiar with relational databases.
- Relational database management systems serve as a useful point of reference for discussing and comparing other non-relational systems.
- Our undergraduate databases course also includes data modelling using the Entity-Relationship (E-R) data model, and it is useful that you are familiar with this before taking this course.

Deeper into relational database management systems

- Very simple model
- Familiar tabular structure
- Has a good theoretical foundation from mathematics (set theory)
- Industrial strength implementations, e.g.
  - Oracle, Sybase, MySQL, PostgreSQL, Microsoft SQL Server, DB2 (IBM mainframes)
- Large user community

Genealogy of Relational Database Management Systems



Key to lines and symbols

DBMS name (Company) | *Acquisition* | v9, 2006 Versions | Discontinued | Branch (intellectual and/or code) | Crossing lines have no special semantics

Felix Naumann, Jana Bauckmann, Claudia Exeler, Jan-Peer Rudolph, Fabian Tschirschnitz
Contact - Hasso Plattner Institut, University of Potsdam, felix.naumann@hpi.de
Design - Alexander Sandt Grafik-Design, Hamburg

Version 6.0 - October 2018
https://hpi.de/naumann/projects/rdbms-genealogy.html

Turing Award Winners: Databases

Charles W. Bachman (1973)

- For his outstanding contributions to database technology

Edgar F. Codd (1981)

- For his fundamental and continuing contributions to the theory and practice of database management systems.

James N. Gray (1998)

- For seminal contributions to database and transaction processing research and technical leadership in system implementation.

Michael Stonebraker (2014)

- For fundamental contributions to the concepts and practices underlying modern database systems.

TDA357/DIT621 course book

- "Database Systems: The Complete Book, 2E", by Hector Garcia-Molina, Jeffrey D. Ullman, and Jennifer Widom
- In the undergraduate Databases course, we mainly look at the “first half” of the textbook (modelling and programming for relational databases and semi-structured data).
- The undergraduate Databases course doesn’t address “Database system implementation” (part IV of the textbook, almost 500 pages).
- This is reasonable, because:
  - many people will **use** database management systems
  - far fewer will **implement** database management systems.

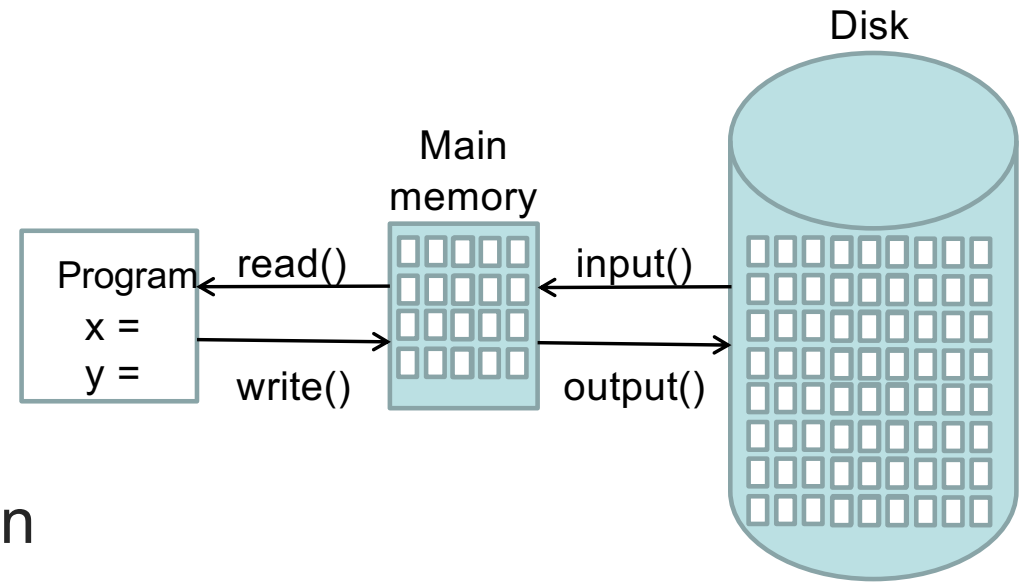
Why look at database system implementation in the new course?

- This course introduces several non-relational technologies (e.g. different flavours of NoSQL system).
- To motivate why such systems have emerged, and to appreciate their advantages and disadvantages, we need deeper knowledge about database technology to understand how these system differ from relational database management systems, and from each other.

- Database system implementation is a worthwhile topic for study in its own right. A full course (or more) could be devoted to this.

Some database system implementation topics

- Database system architecture
- Block buffer
- Transactions and concurrency
- Recovery
- Indexes
- Query processing and optimisation



I saw relational algebra in an undergraduate course – is it actually useful or important?

Yes!

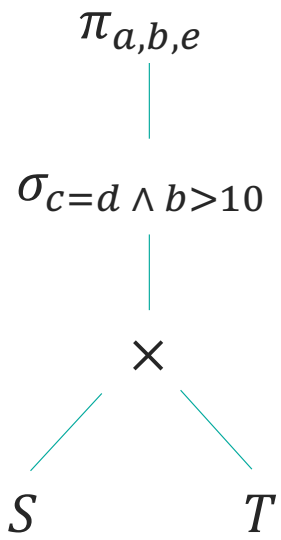
Undergraduate database courses often introduce relational algebra, but sometimes don't have time to explain its usefulness.

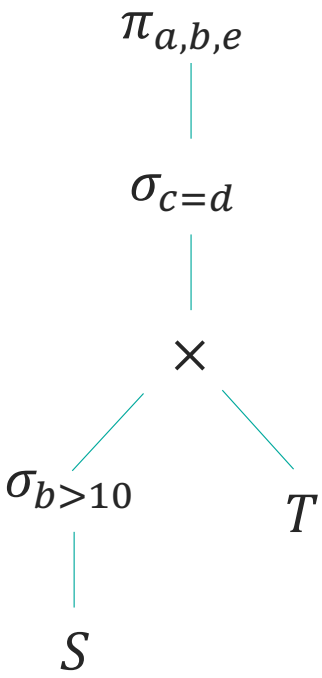
We'll see much more relational algebra in the Advanced Databases course when we look at query processing and optimisation.

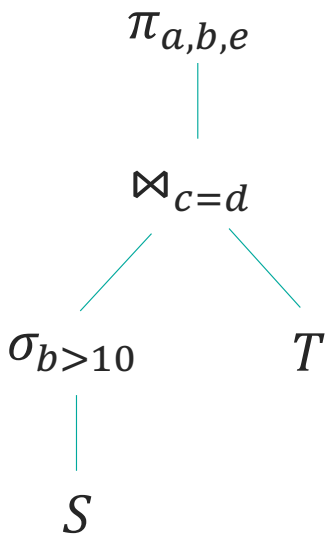
```
SELECT a,b,e
FROM   S,T
WHERE  c=d AND b>10
```

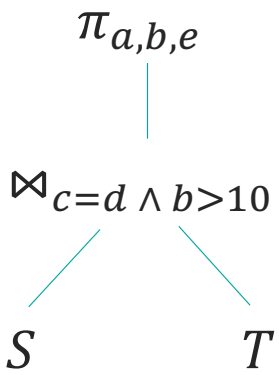
Logical query plans

Suppose we have relations S(a,b,c) and T(d,e).



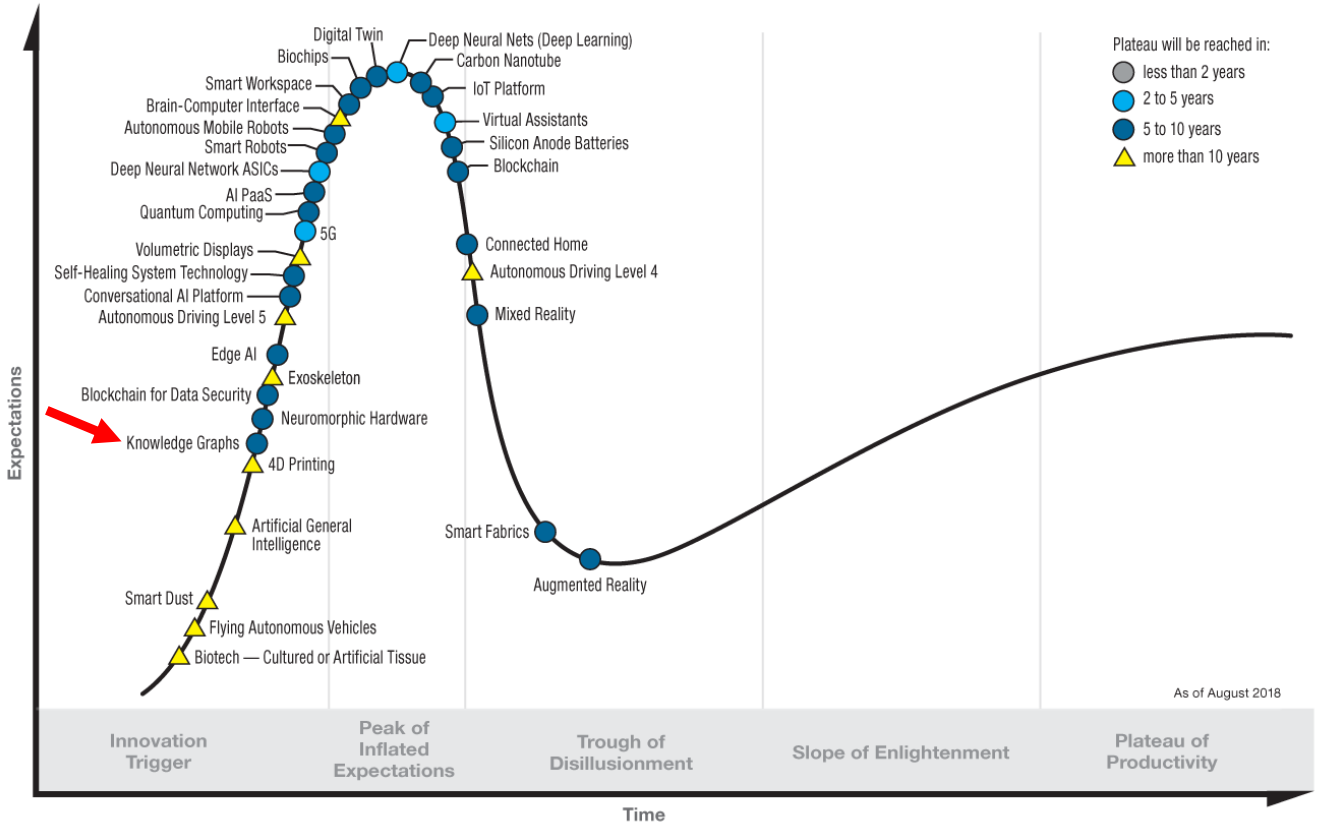






Knowledge Graphs and ontologies

Hype Cycle for Emerging Technologies, 2018



gartner.com/SmarterWithGartner

Source: Gartner (August 2018)

Gartner

Gartner Hype Cycle for Emerging Technologies, 2019

gartner.com/SmarterWithGartner

Source: Gartner

Gartner

Hype Cycle for Emerging Technologies, 2020

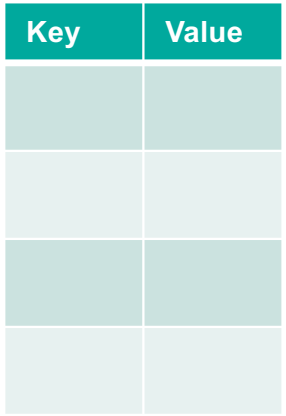
gartner.com/SmarterWithGartner

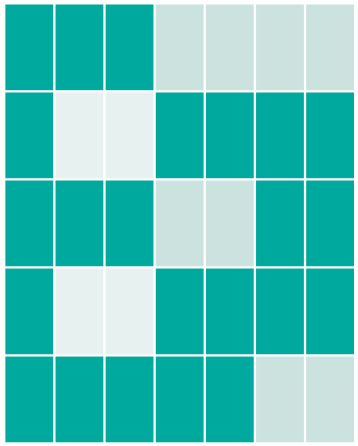
Gartner

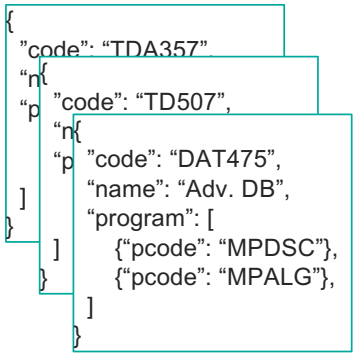
NoSQL

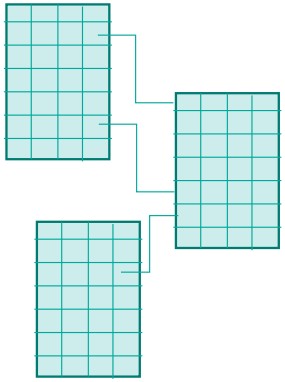
Complex data – which kind of DBMS?

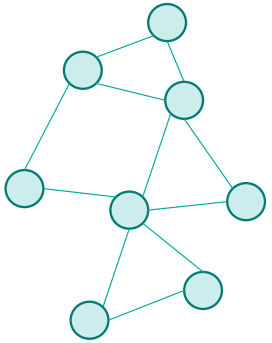
Key-Value | Wide Column | Document | Relational | Graph











Complexity

Graph Databases

GRAPH TECHNOLOGY LANDSCAPE 2020

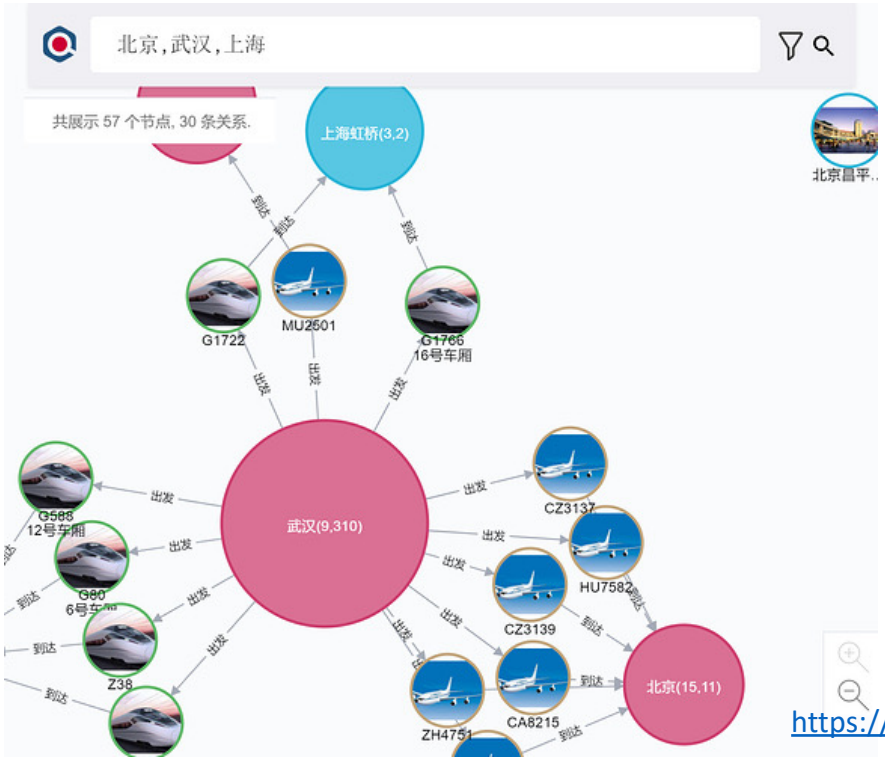
Examples of graph databases

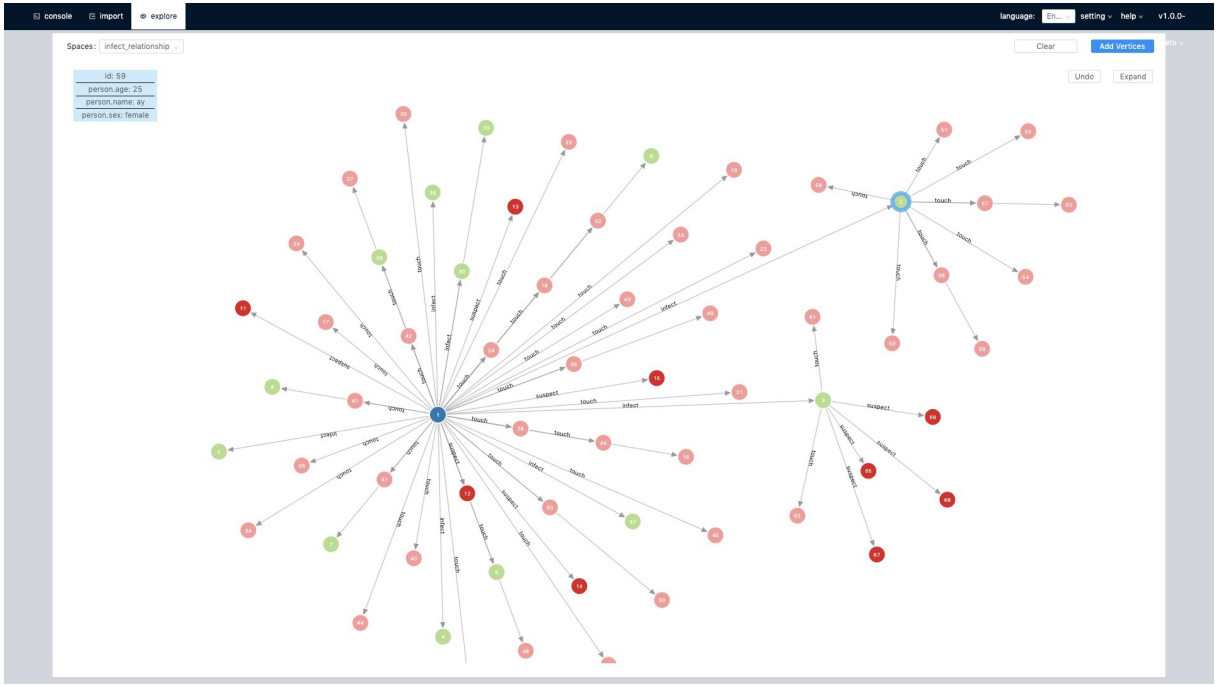
- RDF triples
- SPARQL query language

- Labelled property graph model
- (open)Cypher query language

Database Applications

Contact tracing





https://nebula-graph.io/posts/detect-corona-virus-spreading-with-graph-database/ (2020-02-06)

https://community.neo4j.com/t/fighting-fatal-coronavirus-using-knowledge-graph/14634 (2020-02-12)

Advanced databases

DAT475 | DIT930

Welcome!

UNIVERSITY OF
GOTHENBURG

CHALMERS
UNIVERSITY OF TECHNOLOGY