Introduction

COMPSCI 2DB3: Databases

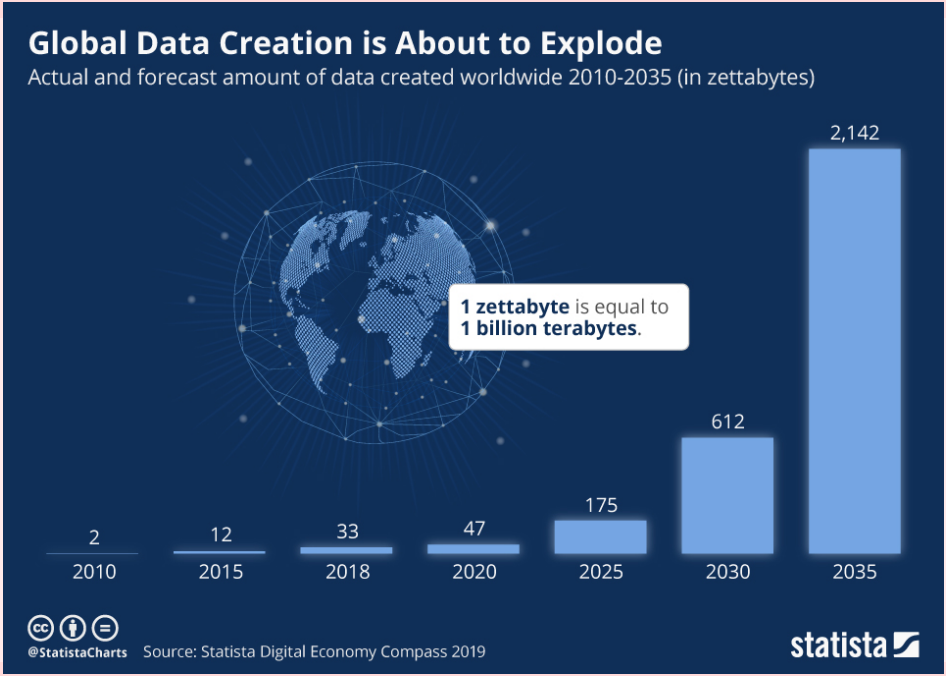
Jelle Hellings Holly Koponen

Department of Computing and Software
McMaster University

Winter 2024

Objectives

- What is data?
- What is a DBMS?
- Scope of this Course.



Source: https://www.statista.com/.

A DAY IN DATA

The exponential growth of data is undisputed, but the numbers behind this explosion – fuelled by internet of things and the use of connected devcies – are hard to comprehend, particularly when looked at in the context of one day

500m

tweets are sent every day

Twitter

4PB

of data created by Facebook, including

350m photos

100m hours of video watch time

Facebook Research

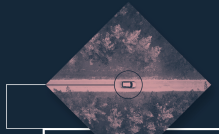
294bn

billion emails are sent

Radicati Group

320bn emails to be sent each day by 2021

306bn emails to be sent each day by 2020



3.9bn

people use emails

4TB

of data produced by a connected car

Intel

DEMYSTIFIYING DATA UNITS

From the more familiar 'bit' or 'megabyte', larger units of measurement are more frequently being used to explain the masses of data

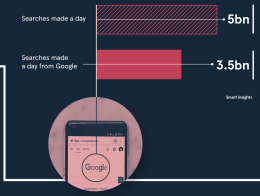
Unit		Value	Size
b	bit	0 or 1	1/8 of a byte
B	byte	8 bits	1 byte
KB	kilobyte	1,000 bytes	1,000 bytes
MB	megabyte	$1,000^2$ bytes	1,000,000 bytes
GB	gigabyte	$1,000^3$ bytes	1,000,000,000 bytes
TB	terabyte	$1,000^4$ bytes	1,000,000,000,000 bytes
PB	petabyte	$1,000^5$ bytes	1,000,000,000,000,000 bytes
EB	exabyte	$1,000^6$ bytes	1,000,000,000,000,000,000 bytes
ZB	zettabyte	$1,000^7$ bytes	1,000,000,000,000,000,000,000 bytes
YB	yottabyte	$1,000^8$ bytes	1,000,000,000,000,000,000,000,000 bytes

*A lowercase "b" is used as an abbreviation for bits, while an uppercase "B" represents bytes.

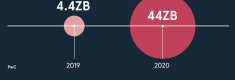
65bn

messages sent over WhatsApp and two billion minutes of voice and video calls made

Facebook



ACCUMULATED DIGITAL UNIVERSE OF DATA



IDC

463EB

of data will be created every day by 2025

IDC

95m

photos and videos are shared on Instagram

Instagram Business

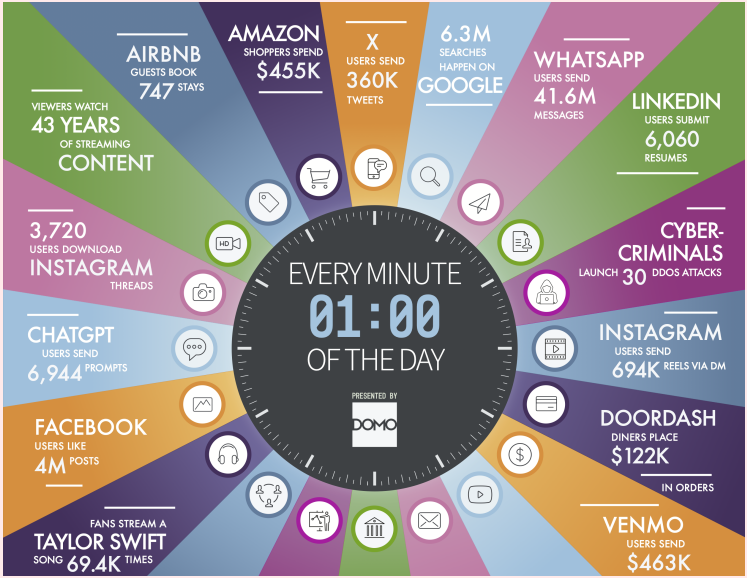
28PB

to be generated from wearable devices by 2020

Statista

RACONTEUR

Source: https://www.raconteur.net/.



Source: https://www.domo.com/.

What is data?

```
Zmqofm49ctWJpryAB31etnDISqToQuSxohTvcPa9vnzew6Qmv
VepIXycFvyEQRndzv9z3MuT296zRcxsXH17tSiSfWcWg8uL9i
DcHE437tJ75Hf94Q35TZc6DcP1eUUBA4fqZgBQc9WeMTSNWlU
luB9ogODArW4wjNOI23XvTsuTP6wyMgQddMEJ4JA3sfWnftSD
```

Question: Is this data?

Vote at https://strawpoll.com/74udy2fzk.
Or: go to https://strawpoll.live and use the code **774380**.

What is data?

```
Zmqofm49ctWJpryAB31etnDISqToQuSxohTvcPa9vnzew6Qmv
VepIXycFvyEQRndzv9z3MuT296zRcxsXH17tSiSfWcWg8uL9i
DcHE437tJ75Hf94Q35TZc6DcP1eUUBA4fqZgBQc9WeMTSNWlU
luB9ogODArW4wjNOI23XvTsuTP6wyMgQddMEJ4JA3sfWnftSD
```

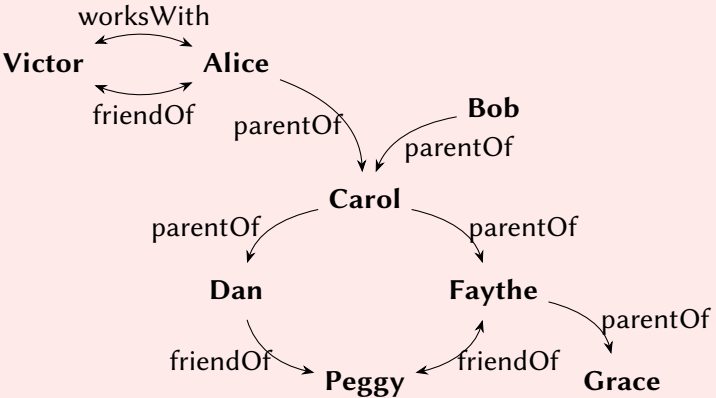
Question: Is this data?

Vote at https://strawpoll.com/74udy2fzk.
Or: go to https://strawpoll.live and use the code **774380**.

Need for structured data

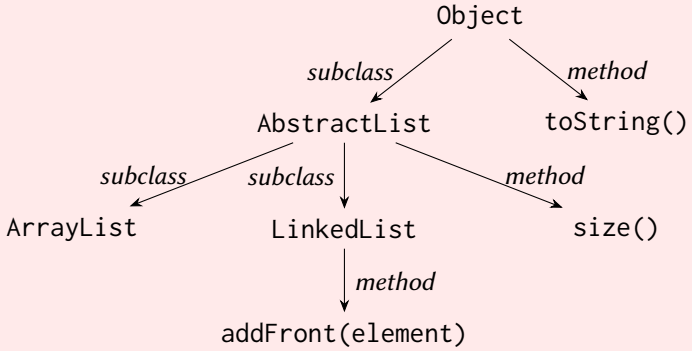
The more we know of the data, the easier it is to work with!

Graph data: RDF, …



Source: https://jhellings.nl/files/lsgda2020_slides.pdf.

Hierarchical data (Trees): XML, JSON, …



Source: https://doi.org/10.1016/j.is.2019.101467.

Structured plain-text data: CSV

Excerpt from the Unicode Character Database

```
005A;LATIN CAPITAL LETTER Z;Lu;0;L;;;;;N;;;;007A;
005B;LEFT SQUARE BRACKET;Ps;0;ON;;;;;Y;OPENING SQUARE BRACKET;;;;
005C;REVERSE SOLIDUS;Po;0;ON;;;;;N;BACKSLASH;;;;
005D;RIGHT SQUARE BRACKET;Pe;0;ON;;;;;Y;CLOSING SQUARE BRACKET;;;;
005E;CIRCUMFLEX ACCENT;Sk;0;ON;;;;;N;SPACING CIRCUMFLEX;;;;
005F;LOW LINE;Pc;0;ON;;;;;N;SPACING UNDERSCORE;;;;
0060;GRAVE ACCENT;Sk;0;ON;;;;;N;SPACING GRAVE;;;;
```

Source: https://www.unicode.org/ucd/.

Table

id	name	program	age	score
53666	Jones	cs	18	3.4
53688	Smith	ee	18	3.2
53650	Smith	math	19	3.8
53831	Madayan	music	11	1.8
53832	Guldu	music	12	2.0

Using data

RDF, XML, JSON, CSV are file formats!

Plenty good libraries in your favorite programming language to deal with them!

Using data

RDF, XML, JSON, CSV are file formats!

Plenty good libraries in your favorite programming language to deal with them!

- How to get ***information*** from the data?
- How to *update* the data?
- How to keep data *consistent*?
- How to deal with *huge amounts* of data?
- How to deal with *concurrent* access to the data?

Using data

RDF, XML, JSON, CSV are file formats!

Plenty good libraries in your favorite programming language to deal with them!

- How to get ***information*** from the data?
- How to *update* the data?
- How to keep data *consistent*?
- How to deal with *huge amounts* of data?
- How to deal with *concurrent* access to the data?

A lot of code even for basic operations—not user friendly or maintainable.

Terminology: Database and DBMS

Database A (very large) collection of data.
Typically all related and with a clear structure.
E.g., University administration:

- *Entities*: students, faculty, courses, and classrooms.
- *Relationships* between entities: course enrollment, lecturer, schedule.

Data model The rules by which real-world data can be represented and structured.

DBMS *DataBase Management System*.
Software designed to assist in maintaining and utilizing databases.

Examples of DBMSs?

- Files & File systems.
- Data analysis tools.
  Microsoft Excel, R, ...
- Websites.

Question: Which ones are DBMSs?

Vote at https://strawpoll.com/6d7yvq6u2.
Or: go to https://strawpoll.live and use the code **597722**.

Examples of DBMSs?

- Files & File systems: ✖.
- Data analysis tools: ✖.
  Microsoft Excel, R, ...
- Websites: ✖.
  Most websites use a DBMS for their data needs.

Question: Which ones are DBMSs?

Vote at https://strawpoll.com/6d7yvq6u2.
Or: go to https://strawpoll.live and use the code **597722**.

Is a file system a DBMS?

Question: What happens in the following situation?

- Two users: you and a project partner.
- A single file *a*: *important_data.txt*.
- Both of you are editing *a* via a text editor.
- Both of you store at the same time.

Vote at https://strawpoll.com/9dd7dh3w5.
Or: go to https://strawpoll.live and use the code **197566**.

Is a file system a DBMS?

Question: What happens in the following situation?

- Two users: you and a project partner.
- A single file *a*: *important_data.txt*.
- Both of you are editing *a* via a text editor.
- Both of you store at the same time.

Vote at `https://strawpoll.com/9dd7dh3w5`.
Or: go to `https://strawpoll.live` and use the code **197566**.

Anything can happen

Details depend on editor(s), file system, operating system, hardware, ...

From files to DBMS

What happens when writing a file data during a power surge?

Details depend on editor(s), file system, operating system, hardware, ...
Nowadays file systems that do something sensible exist.

How to make data management reliable?

- Manual: very tricky and rather complex.
- Use a DBMS: easy-to-use abstraction with strong guarantees.

Main benefits of a DBMS

Data Independence

How you *use* the data is independent of how the data is *represented* and *stored*.

High-level query language

Specify *what* information you want, not *how to get it*.

Main benefits of a DBMS

Data Independence

How you *use* the data is independent of how the data is *represented* and *stored*.

High-level query language

Specify *what* information you want, not *how to get it*.

DBMSs provide

- strong guarantees: data integrity and recovery;
- concurrent data access;
- efficiency even with large data volumes (using external storage);
- ....

Why should I care?

Data drives modern society and economy

Practitioner Strong industry demand for DBMS specialists that set up databases.

Developer A lot of software deals with data.
E.g., websites, administrative systems, IoT sensor networks, ...

Research Database research and engineering encompasses computer science.
E.g., algorithms, operating systems, formal languages, machine learning, ...
Research with strong motivations from applications and industry.

Our focus: The relational data model

Simplified description: Data represented by a bunch of tables

- *Versatile*: can model most real-world data effectively and efficiently.
- *Widely used*: most real-world systems (partly) rely on relational databases (e.g., online forums, online blogs, university administration, sales records, …)
- Centered around *combining data* (via queries) to obtain new information.
- *Complex enough* to model real-world data and derive complex new information.
- *Simple enough* to access and query data in a user-friendly (high-level) manner *while* abstracting away from complex high-performance implementation details.

Our focus: The relational data model

Simplified description: Data represented by a bunch of tables

- *Versatile*: can model most real-world data effectively and efficiently.
- *Widely used*: most real-world systems (partly) rely on relational databases (e.g., online forums, online blogs, university administration, sales records, …)
- Centered around *combining data* (via queries) to obtain new information.
- *Complex enough* to model real-world data and derive complex new information.
- *Simple enough* to access and query data in a user-friendly (high-level) manner *while* abstracting away from complex high-performance implementation details.

Many alternatives: graph databases, key-value stores, document stores, …

- Typically highly-specialized to excel in a specific situation.
- Have sever limitations outside their area of specialization.

What is in this course?

An introduction to databases

We will look at the fundamental aspects of data management.

- How to *model data*: determine the structure of data.
- How to *query data*: interacting with data & derive new information.
- How to *maintain integrity of data*: e.g., via constraints.
- How to reason about the *quality* of the structure of our data.
- How to *efficiently implement* storing and querying data.