Introduction

Introduction to Database Management

CS348 Fall 2022

About the Instructor

- **Xi** He

- Why CS348: intro to database management?

- What is my expectation?

- What else do I (plan to) do besides CS348 ?

Scan the QR code to get the google form & fill it up

https://forms.gle/n89cV76pDxnr5fyd7

More about the Teaching Team

- Instructor: Xi He
  - Email: xihe@uwaterloo.ca
  - https://cs.uwaterloo.ca/~xihe/

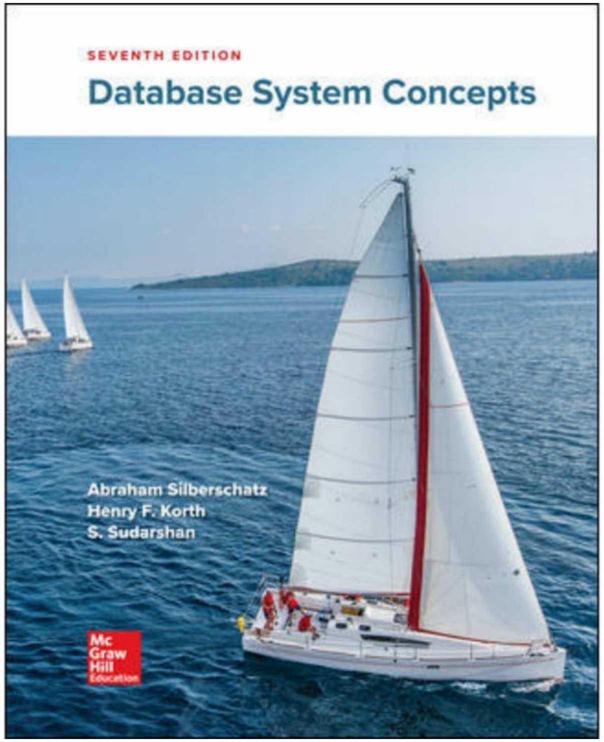
- Instructional support coordinator: Sylvie Davies
  - Email: sldavies@uwaterloo.ca

- IAs and TAs:
  - Guy Coccimiglio
  - Glaucia Melo
  - Krishna Kanth Arumugam
  - Partha Chakraborty
  - Chang Liu
  - Nimmi Rashinika Weeraddana
  - Ruoxi Zhang

- Office hours will be posted on Learn/Piazza

Textbook

- Database System Concepts (Seventh Edition) Abraham Silberschatz, Henry F. Forth and S.Sudarshan, McGraw Hill.



Logistics

- Course Website:
  - https://student.cs.uwaterloo.ca/~cs348/
  - Course schedule, lecture notes
- Learn:
  - https://learn.uwaterloo.ca/
  - Assignment questions/partial solutions, project info
- Piazza for student discussion, Q&A, TAs info:
  - piazza.com/uwaterloo.ca/fall2022/cs348/home
  - Q&A: 24 hours policy
- Work submission: Crowdmark/Marmoset/Learn
  - Watch your emails for the links

Marking and Late Policies

- Marking and appeals:
  - For everything, there will be an appeal deadline that will be indicated on the front page
  - No appeals will be accepted past this date unless you were sick the entire period until the appeal date
- Late assignments
  - Late assignments will be accepted for 48 hours past the due date, but…
  - For each 24 hour past the due date, a 5% penalty will be applied (cumulatively) off the top

Community Standard

- Group discussion for assignments is okay (and encouraged), but
  - Acknowledge any help you receive from others
  - Make sure you “own” your solution
- All suspected cases of violation will be aggressively pursued

Assessments

- 3 Assignments
- 1 Midterm Exam (Nov 4, Fri)
- 1 Final Exam (TBD)
- Group Project (Optional): Only 1 mark breakdown is allowed

Mark Breakdown	Project-based	Exam-based
3 Assignments	30%	30%
Midterm Exam	10%	30%
Final Exam	20%	40%
Project	40%	-

- Examples:
  - Project based: 90 pts (A/P) * 70% + 70 pts (E) * 30% = 84 pts
  - Exam based: 90 pts (A) * 30% + 80 pts (E) * 70 = 83 pts

Lectures

- Lecture slides released on Course Website before Tue/Thur 00:01am

- Lecture format:
  - Important announcements (Don’t miss this!)
  - Key points and examples
  - Exercises with partial solutions

Project

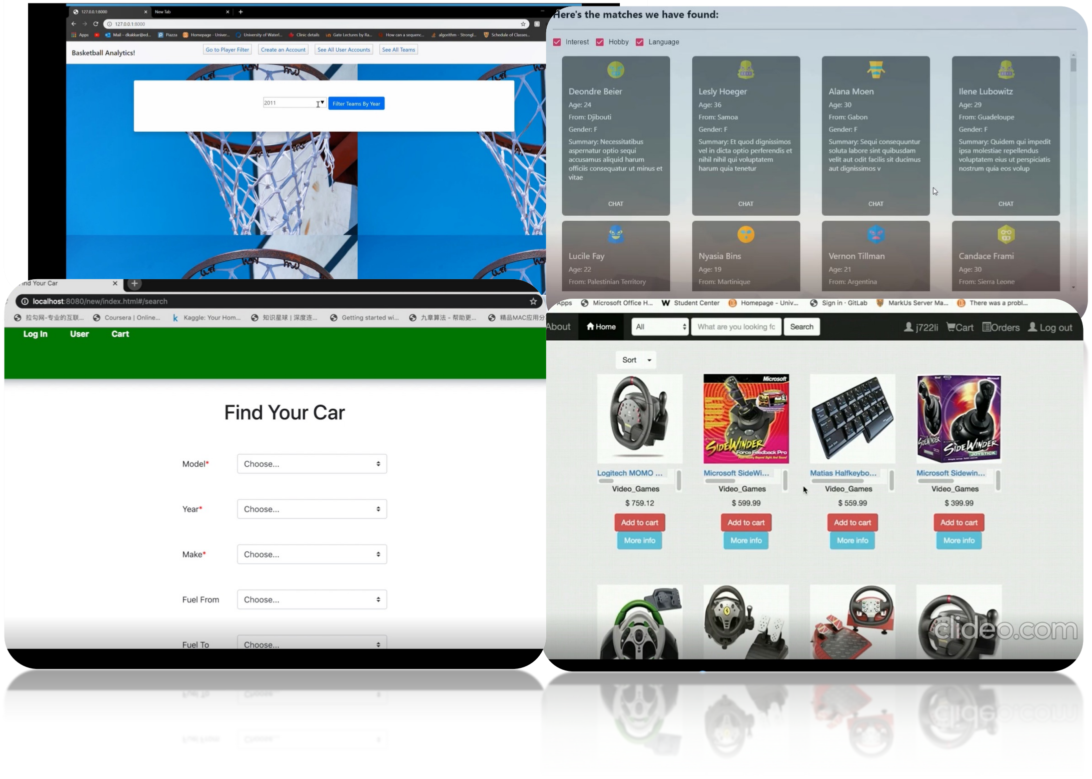
- Team of 4-5 students
- DB-supported applications
- Project timeline
  - Milestone 0 (form a team) week 4 (Sep 27, Tue)
  - Milestone 1 (proposal) week 7 (Oct 20, Thur)
  - Milestone 2 (mid-term report) week 11 (Nov 17, Thur)
  - Final (report + demo) week 13 (Dec 1, Thur)

- More details will be released in week 2, but you can start to brainstorm and find your teammates!
  - Members from different sessions are allowed.
  - Piazza is a good place to find teammates.

Project

- Project demos from previous years

Video Demo for NBA Season Statistics



Basketball Analytics!

Go to Player Filter | Create an Account | See All User Accounts | See All Teams

2011 | Filter Teams By Year

Here's the matches we have found:

☑ Interest ☑ Hobby ☑ Language

Deondre Beier
Age: 24
From: Djibouti
Gender: F
Summary: Necessitatibus aspernatur optio sequi accusamus aliquid harum officiis consequatur ut minus et vitae
CHAT

Lesly Hoeger
Age: 36
From: Samoa
Gender: F
Summary: Et quod dignissimos vel in dicta optio perferendis et nihil nihil qui voluptatem harum quia tenetur
CHAT

Alana Moen
Age: 30
From: Gabon
Gender: F
Summary: Sequi consequuntur soluta labore sint quibusdam velit aut odit facilis sit ducimus aut dignissimos v
CHAT

Ilene Lubowitz
Age: 29
From: Guadeloupe
Gender: F
Summary: Quidem qui impedit ipsa molestiae repellendus voluptatem eius ut perspiciatis nostrum quia eos volup
CHAT

Lucile Fay
Age: 22
From: Palestinian Territory

Nyasia Bins
Age: 19
From: Martinique

Vernon Tillman
Age: 21
From: Argentina

Candace Frami
Age: 30
From: Sierra Leone

localhost:8080/new/index.html#/search

Log In | User | Cart

Find Your Car

Model* Choose...

Year* Choose...

Make* Choose...

Fuel From Choose...

Fuel To Choose...

About | Home | All | What are you looking fc | Search | j722li | Cart | Orders | Log out

Sort

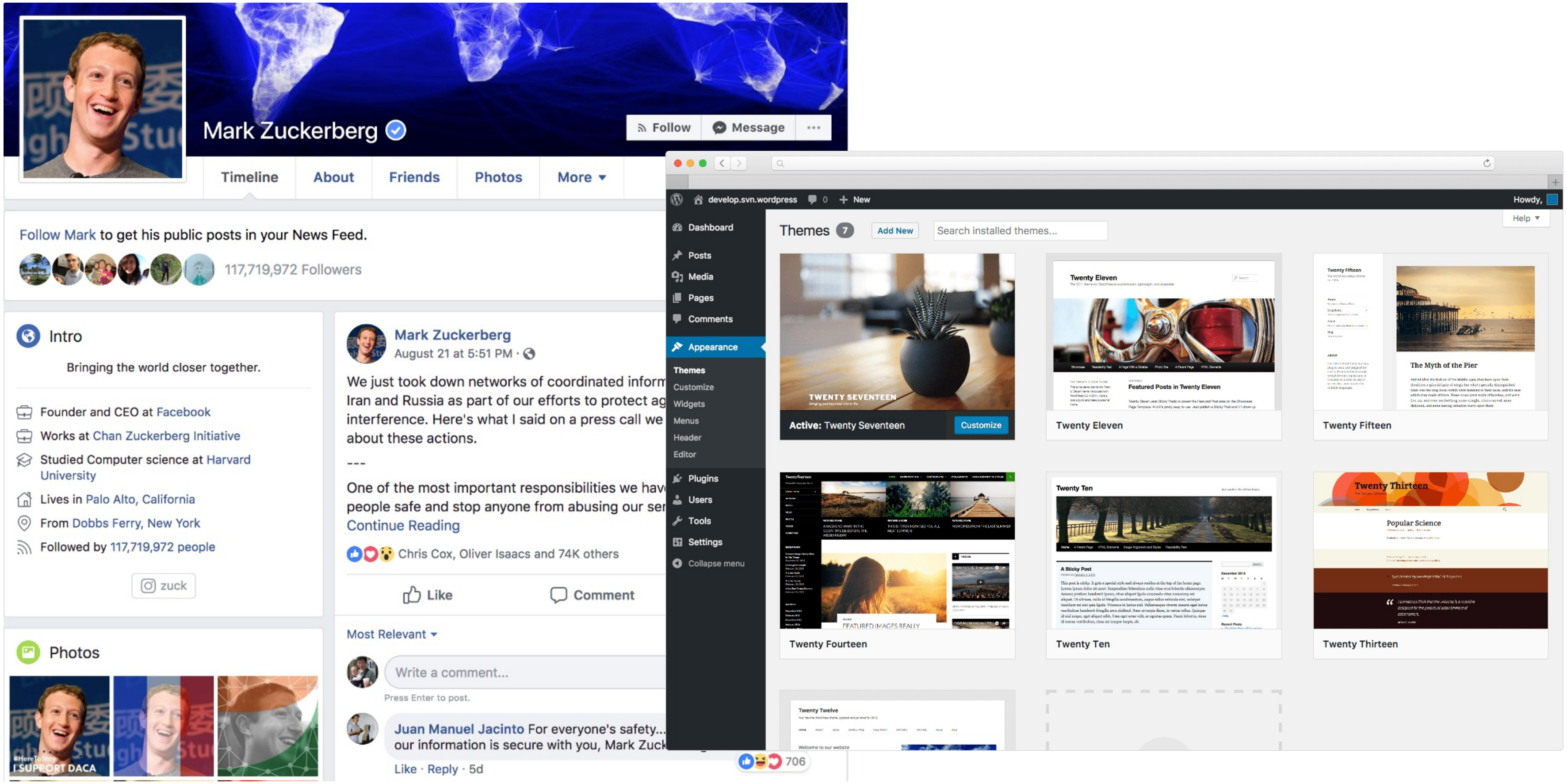
Logitech MOMO ...	Microsoft SideWi...	Matias Halfkeybo...	Microsoft Sidewin...
Video_Games	Video_Games	Video_Games	Video_Games
$ 759.12	$ 599.99	$ 559.99	$ 399.99
Add to cart	Add to cart	Add to cart	Add to cart
More info	More info	More info	More info

Your turn to be creative

http://www.yummymummyclub.ca/sites/default/files/styles/large/public/field/image/teaching_kids_creative_skills.jpg

What comes to your mind when you think about “databases”?

Assignment 1 – warm up Qns



Mark Zuckerberg

Follow · Message · ···

Timeline · About · Friends · Photos · More ▾

Follow Mark to get his public posts in your News Feed.

117,719,972 Followers

Intro

Bringing the world closer together.

- Founder and CEO at Facebook
- Works at Chan Zuckerberg Initiative
- Studied Computer science at Harvard University
- Lives in Palo Alto, California
- From Dobbs Ferry, New York
- Followed by 117,719,972 people

zuck

Photos

Mark Zuckerberg
August 21 at 5:51 PM ·

We just took down networks of coordinated inform
Iran and Russia as part of our efforts to protect ag
interference. Here's what I said on a press call we
about these actions.

One of the most important responsibilities we hav
people safe and stop anyone from abusing our ser
Continue Reading

Chris Cox, Oliver Isaacs and 74K others

Like · Comment

Most Relevant ▾

Write a comment...

Press Enter to post.

Juan Manuel Jacinto For everyone's safety...
our information is secure with you, Mark Zuck

Like · Reply · 5d

706

develop.svn.wordpress · 0 · + New · Howdy,

Dashboard · Posts · Media · Pages · Comments · Appearance · Themes · Customize · Widgets · Menus · Header · Editor · Plugins · Users · Tools · Settings · Collapse menu

Themes 7 · Add New · Search installed themes... · Help ▾

Active: Twenty Seventeen · Customize

Twenty Eleven

Twenty Fifteen

Twenty Fourteen

Twenty Ten

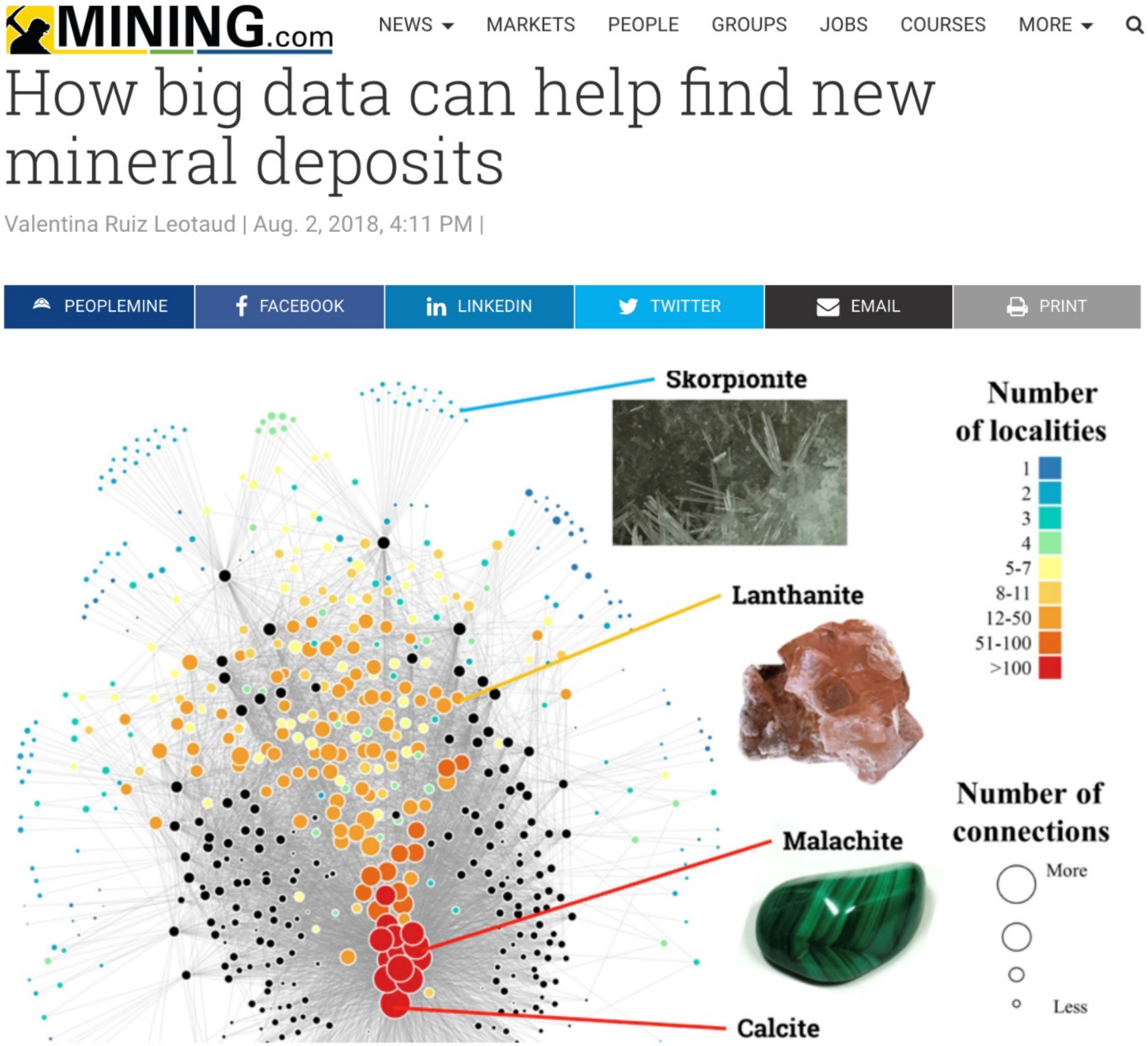
Twenty Thirteen

Twenty Twelve

Facebook uses ? to store posts

WordPress uses ? to manage components of a website (pages, links, menus, etc.)

Data → gold

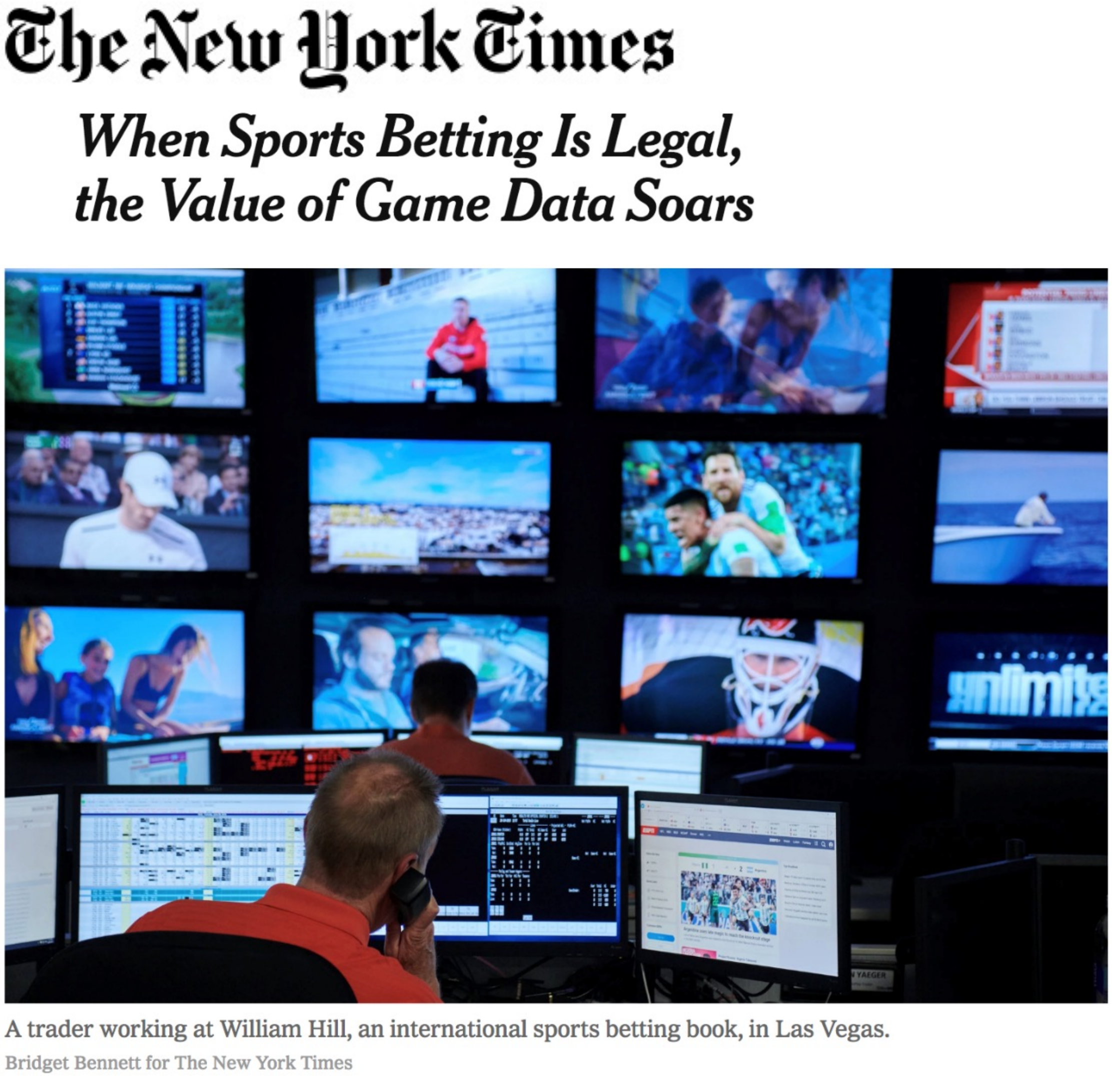


How big data can help find new mineral deposits

Valentina Ruiz Leotaud | Aug. 2, 2018, 4:11 PM |

http://www.mining.com/big-data-can-help-find-new-mineral-deposits/

Data → fun and profit



The New York Times

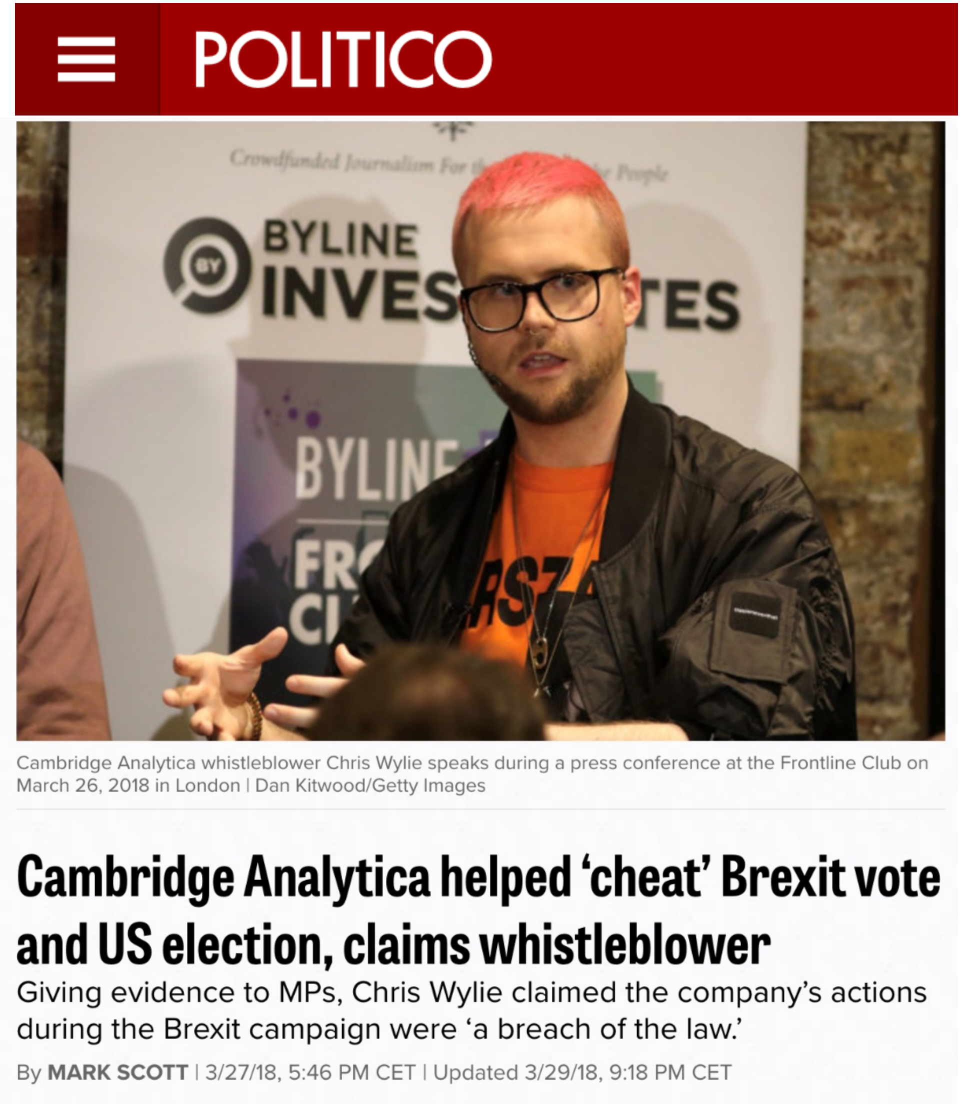
When Sports Betting Is Legal, the Value of Game Data Soars

A trader working at William Hill, an international sports betting book, in Las Vegas.

Bridget Bennett for The New York Times

https://www.nytimes.com/2018/07/02/sports/sports-betting.html

Data → power



POLITICO

Cambridge Analytica whistleblower Chris Wylie speaks during a press conference at the Frontline Club on March 26, 2018 in London | Dan Kitwood/Getty Images

Cambridge Analytica helped 'cheat' Brexit vote and US election, claims whistleblower

Giving evidence to MPs, Chris Wylie claimed the company's actions during the Brexit campaign were 'a breach of the law.'

By **MARK SCOTT** | 3/27/18, 5:46 PM CET | Updated 3/29/18, 9:18 PM CET

https://www.politico.eu/article/cambridge-analytica-chris-wylie-brexit-trump-britain-data-protection-privacy-facebook/

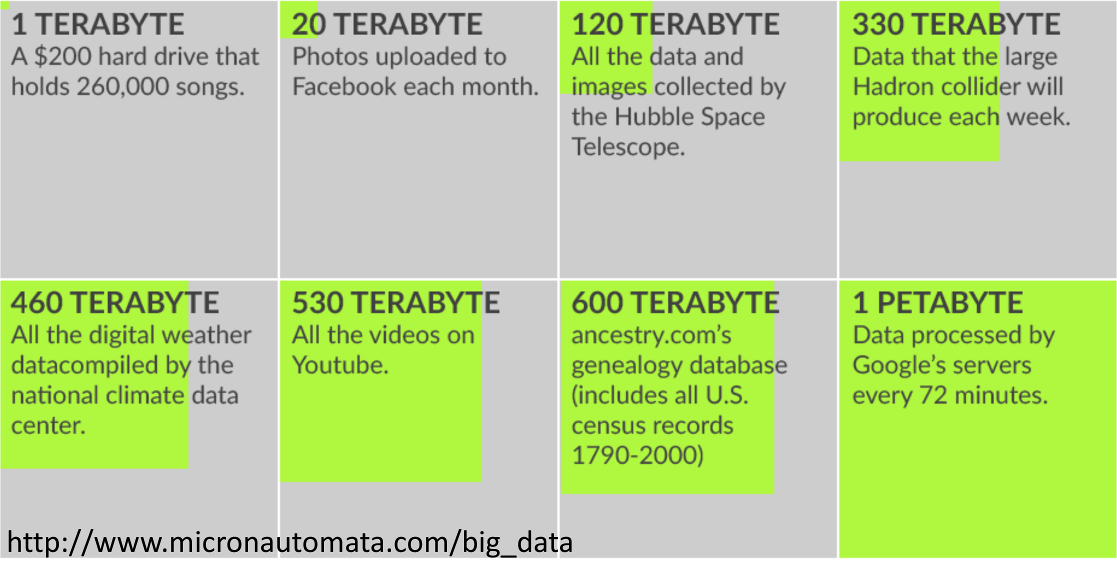
Democratizing data (and analysis)

- Democratization of data: more data—relevant to you and the society—are being collected
  - “Smart planet”
  - “Government in the sunshine”

- But few people know how to analyze them

Challenges

- Moore's Law:
  Processing power doubles every 18 months
- But amount of data doubles every 9 months
  - Disk sales (# of bits) doubles every 9 months
  - Parkinson's Law: *Data expands to fill the space available for storage*



1 TERABYTE
A $200 hard drive that holds 260,000 songs.

20 TERABYTE
Photos uploaded to Facebook each month.

120 TERABYTE
All the data and images collected by the Hubble Space Telescope.

330 TERABYTE
Data that the large Hadron collider will produce each week.

460 TERABYTE
All the digital weather datacompiled by the national climate data center.

530 TERABYTE
All the videos on Youtube.

600 TERABYTE
ancestry.com's genealogy database (includes all U.S. census records 1790-2000)

1 PETABYTE
Data processed by Google's servers every 72 minutes.

http://www.micronautomata.com/big_data

Moore's Law reversed

Time to process all data
doubles every 18 months!

- Does your attention span double every 18 months?
  - No, so we need smarter data management and processing techniques

So, what is a database system?

From Oxford Dictionary:

- Database: an organized body of related information
- Database system, DataBase Management System (DBMS): a software system that facilitates the creation and maintenance and use of an electronic database

What do you want from a DBMS?

Consider bank applications …

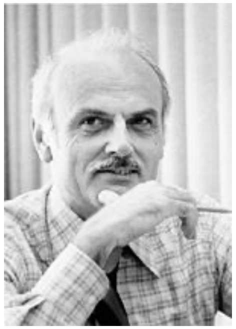
- Data: Each account belongs to a branch, has a number, an owner, a balance, …; each branch has a location, a manager, …
- Persistency: Balance can’t disappear after a power outage
- Query: What’s the balance in Homer Simpson’s account? What’s the difference in average balance between Springfield and Capitol City accounts?
- Modification: Homer withdraws $100; charge accounts with lower than $500 balance a $5 fee

How to design a DBMS?

3 Turing Award Winners!

- Charles Bachman, 1973

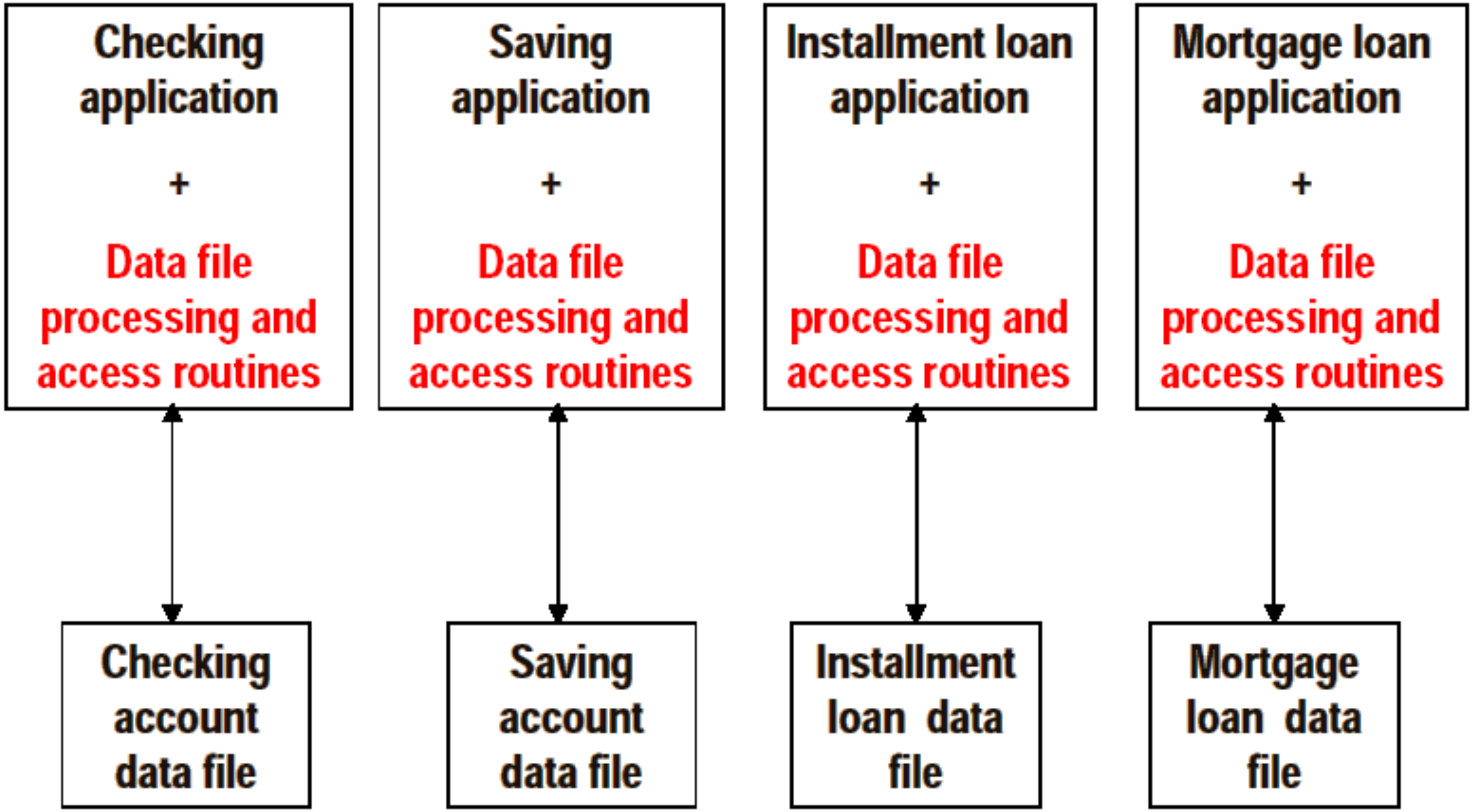
- Edgar F. Codd, 1981



- Michael Stonebraker, 2014

https://amturing.acm.org/award_winners/bachman_9385610.cfm
https://en.wikipedia.org/wiki/Edgar_F._Codd
https://en.wikipedia.org/wiki/Michael_Stonebraker

The birth of DBMS – 1



Checking application + Data file processing and access routines

Saving application + Data file processing and access routines

Installment loan application + Data file processing and access routines

Mortgage loan application + Data file processing and access routines

Checking account data file

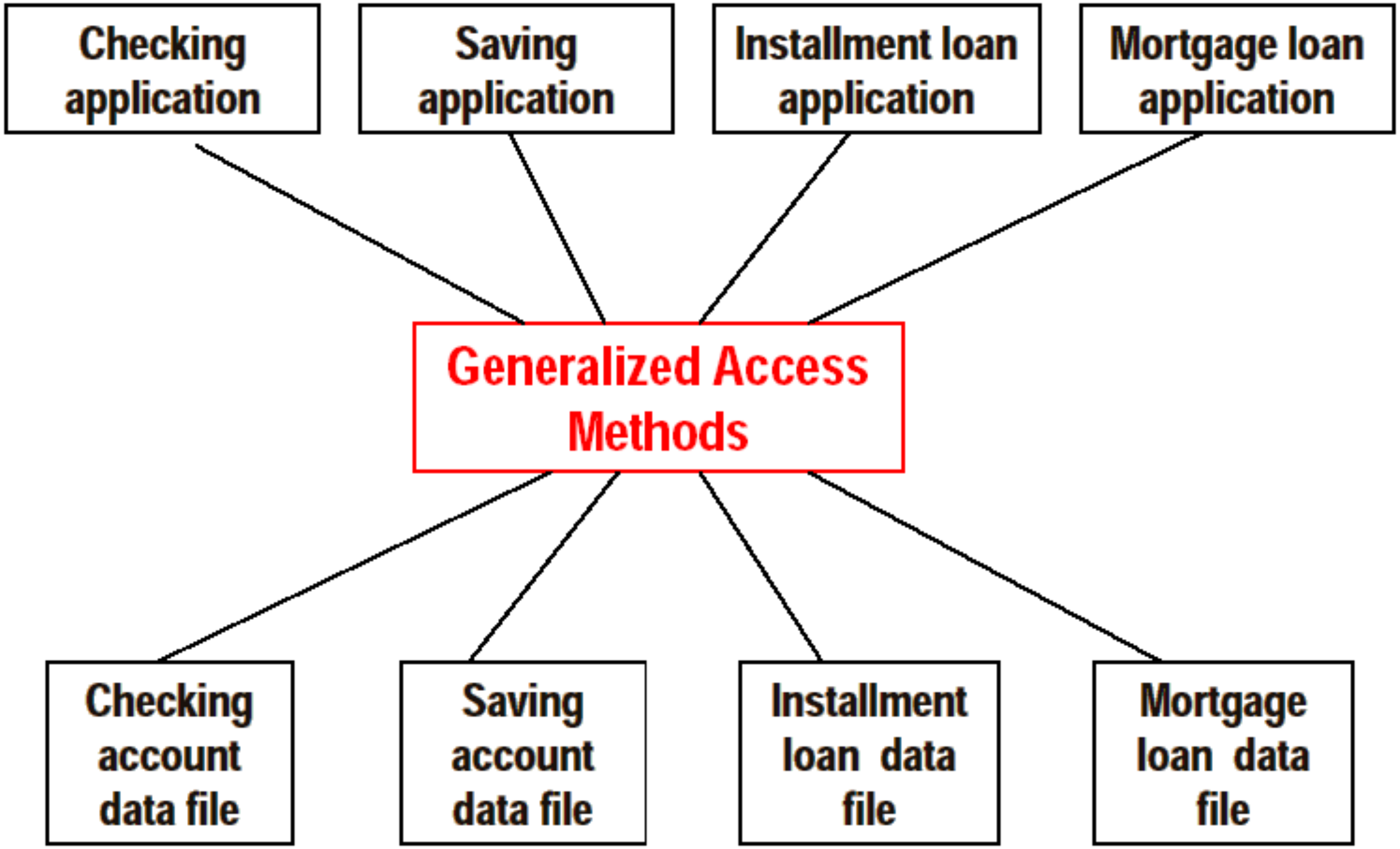
Saving account data file

Installment loan data file

Mortgage loan data file

From Hans-J. Schek's *VLDB* 2000 slides

The birth of DBMS – 2



Checking application

Saving application

Installment loan application

Mortgage loan application

Generalized Access Methods

Checking account data file

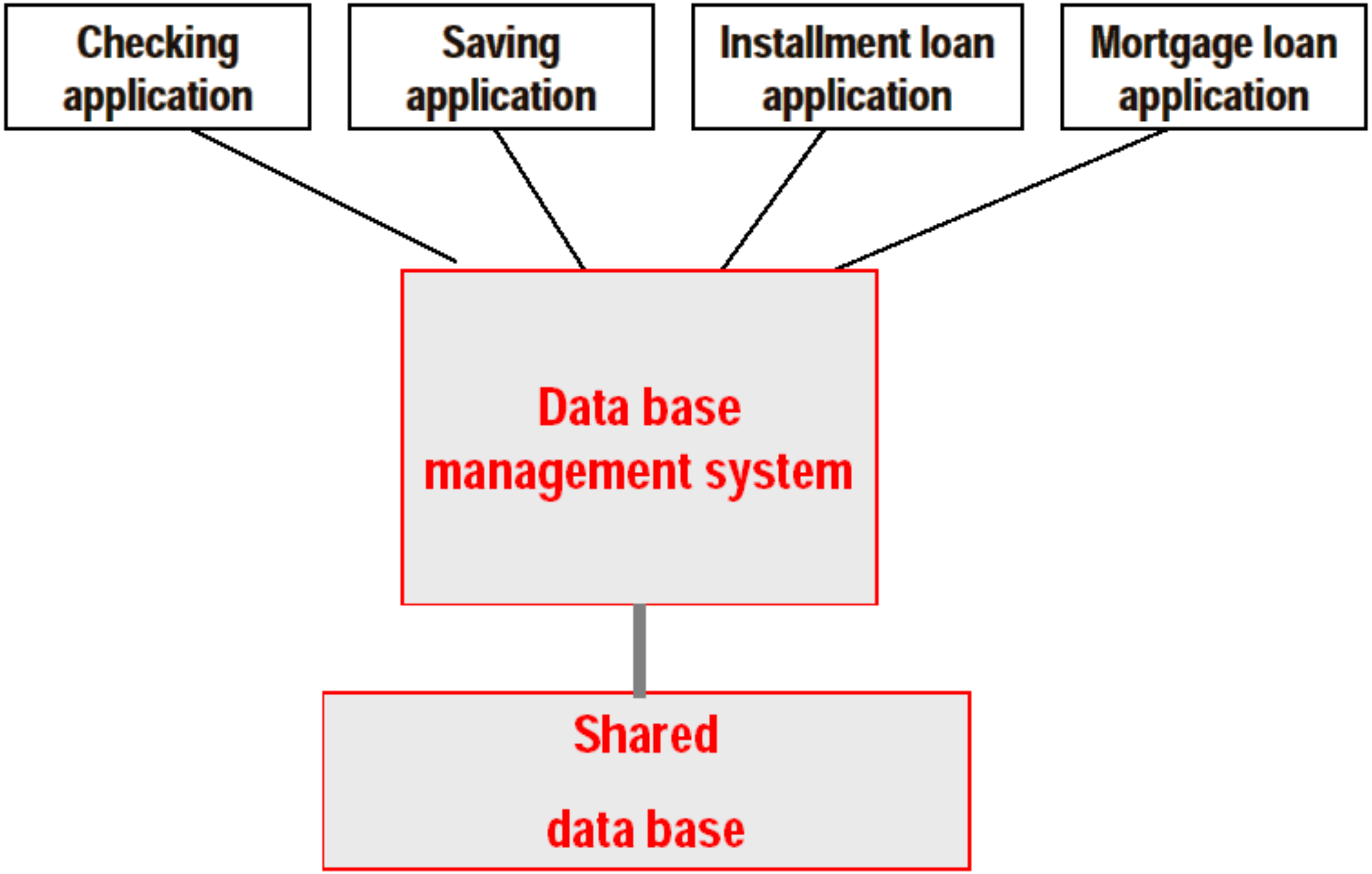
Saving account data file

Installment loan data file

Mortgage loan data file

From Hans-J. Schek's *VLDB* 2000 slides

The birth of DBMS – 3


Checking application
Saving application
Installment loan application
Mortgage loan application
Data base management system
Shared data base

From Hans-J. Schek’s *VLDB* 2000 slides

Early efforts

- "Factoring out" data management functionalities from applications and standardizing these functionalities is an important first step
  - CODASYL standard (circa 1960's)

  ☞ Bachman got a Turing award for this in 1973

- But getting the abstraction right (the API between applications and the DBMS) is still tricky

https://amturing.acm.org/award_winners/bachman_9385610.cfm

Continue with our bank example…

- Query: Who have accounts with 0 balance managed by a branch in Springfield?
- Pseudo-code of a CODASYL application:

```
Use index on account(balance) to get accounts with 0 balance;
For each account record:
        Get the branch id of this account;
        Use index on branch(id) to get the branch record;
        If the branch record’s location field reads “Springfield”:
                Output the owner field of the account record.
```

- Programmer controls “navigation”: accounts → branches
  - How about branches → accounts?

What’s wrong?

With the CODASYL approach, to write correct & efficient code, programmers need to

- know how data is organized physically
- worry about data/workload characteristics

The relational revolution (1970’s)

- A simple model: data is stored in relations (tables)

Account_id	*name*	*balance*	*Branch_id*
142	Bart	10000	2
123	Milhouse	0	1
…	…	…	…

Branch_id	*location*
1	Springfield
2	Summerfield
…	…

- A declarative query language: SQL

```
SELECT Account.owner
        FROM Account, Branch
        WHERE Account.balance = 0
                AND Branch.location = 'Springfield’
                AND Account.branch_id = Branch.branch_id;
```

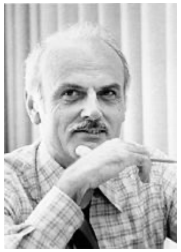
The relational revolution (1970's)

- Programmers specifies what answers a query should return, but not how the query is executed

- DBMS picks the best execution strategy based on physical structure of the data, and data/workload characteristics, etc.

☞Provides physical data independence

- And a Turing Award for E. F. Codd in 1981

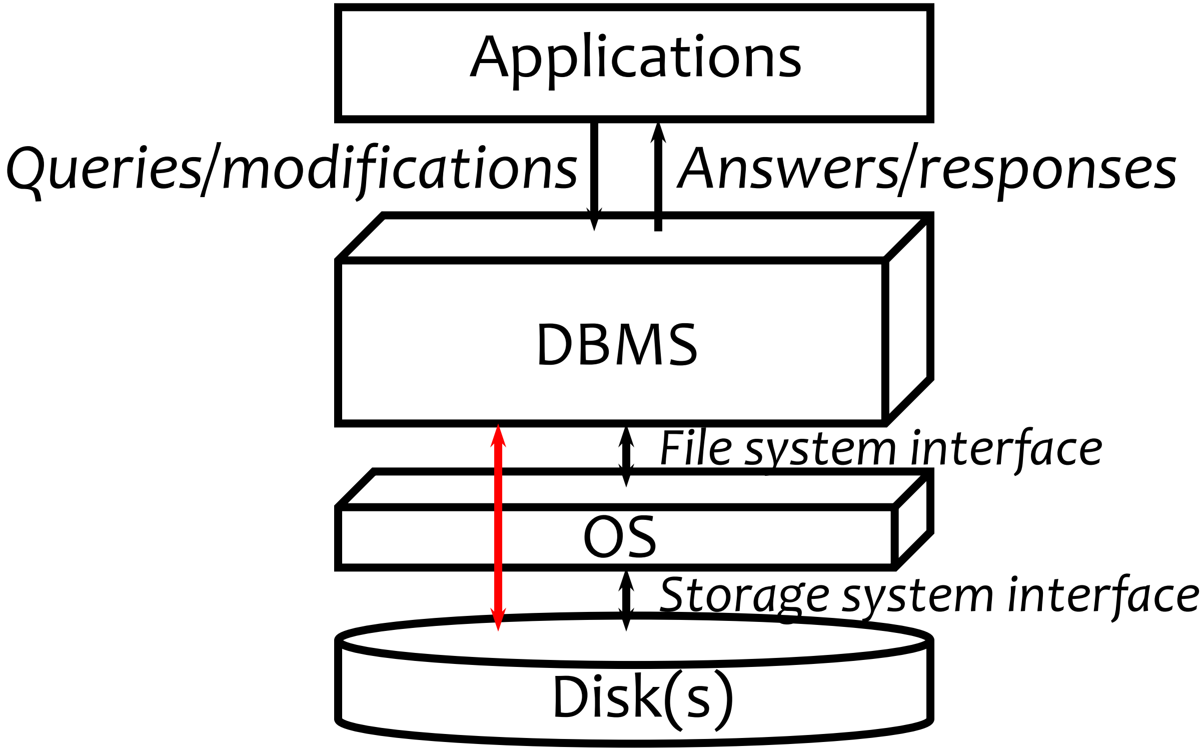


https://en.wikipedia.org/wiki/Edgar_F._Codd

Standard DBMS features

- Logical data model; declarative queries and updates → physical data independence
- Multi-user concurrent access; persistent storage of data; safety from system failures
- Performance, performance, performance

Standard DBMS architecture

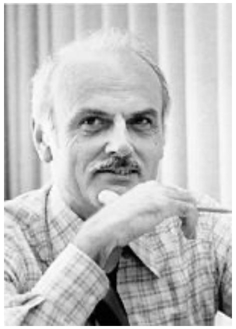


- Much of the OS may be bypassed for performance and safety

Modern DBMSs

- Charles Bachman, 1973

- Edgar F. Codd, 1981



- Michael Stonebraker, 2014

Relational DBMS (e.g. Ingres, Postgres) and modern DBMSs (e.g. C-store, H-store)

https://amturing.acm.org/award_winners/bachman_9385610.cfm
https://en.wikipedia.org/wiki/Edgar_F._Codd
https://en.wikipedia.org/wiki/Michael_Stonebraker

Course components

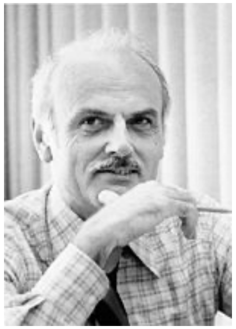
- Relational databases (Lectures 1-10)
  - Relational algebra, SQL, app programming, database design
- Database internals (Lectures 11-15)
  - Storage, indexing, query processing and optimization, transactions
- Advanced topics (Optional)
  - Concurrency & recover, parallel data processing/MapReduce, distributed/parallel dbms, data warehousing and data mining, privacy etc.

Summary

- Charles Bachman, 1973

CODASYL standard

- Edgar F. Codd, 1981



Relational databases
“Physical data independence”

- Michael Stonebraker, 2014

Relational DBMS (e.g. Ingres, Postgres) and modern DBMSs (e.g. C-store, H-store, SciDB)

https://amturing.acm.org/award_winners/bachman_9385610.cfm
https://en.wikipedia.org/wiki/Edgar_F._Codd
https://en.wikipedia.org/wiki/Michael_Stonebraker

What’s next?

- Lecture 2: Relational model and relational algebra