ΠΑΝΕΠΙΣΤΗΜΙΟ ΠΕΙΡΑΙΩΣ
ΤΜΗΜΑ ΠΛΗΡΟΦΟΡΙΚΗΣ

ΠΜΣ Κυβερνοασφαλεια
και Επιστημη Δεδομενων

MSc Cybersecurity
and Data Science

DEPT OF INFORMATICS
UNIVERSITY OF PIRAEUS

Διαχείριση Μεγάλων Δεδομένων

Big Data Management

1 – Εισαγωγή, Θέματα αιχμής στη διαχείριση δεδομένων, Επισκόπηση σχεσιακών ΣΔΒΔ

DATA STORIES
Data Science Lab. @ University of Piraeus

v. 2022.10

Περίγραμμα μαθήματος

- **τι**:
  - Εισαγωγή – Θέματα αιχμής στη Διαχείριση Δεδομένων (ΔΔ) – Επισκόπηση «παραδοσιακών» (Σχεσιακών) ΣΔΒΔ
  - Μη-παραδοσιακή αρχιτεκτονική ΣΔΒΔ (κατανεμημένα, στο υπολογιστικό νέφος – η εποχή των «Μεγάλων Δεδομένων» (“Big Data”)
  - Μη-παραδοσιακή μοντελοποίηση δεδομένων (συστήματα NoSQL)
- **πώς**:
  - Θεωρητικές διαλέξεις (6)
  - Εργαστηριακές διαλέξεις (4) σε MongoDB, Spark Batch/Streaming/MΛιβ
- **ποιοι**:
  - Διδάσκοντες: αναπλ. καθ. Νίκος Πελέκης, δρ. Γιώργος Παπαστεφανάτος
  - Εργαστηριακοί βοηθοί: Γ. Αλεξίου, Σ. Μαρούλης

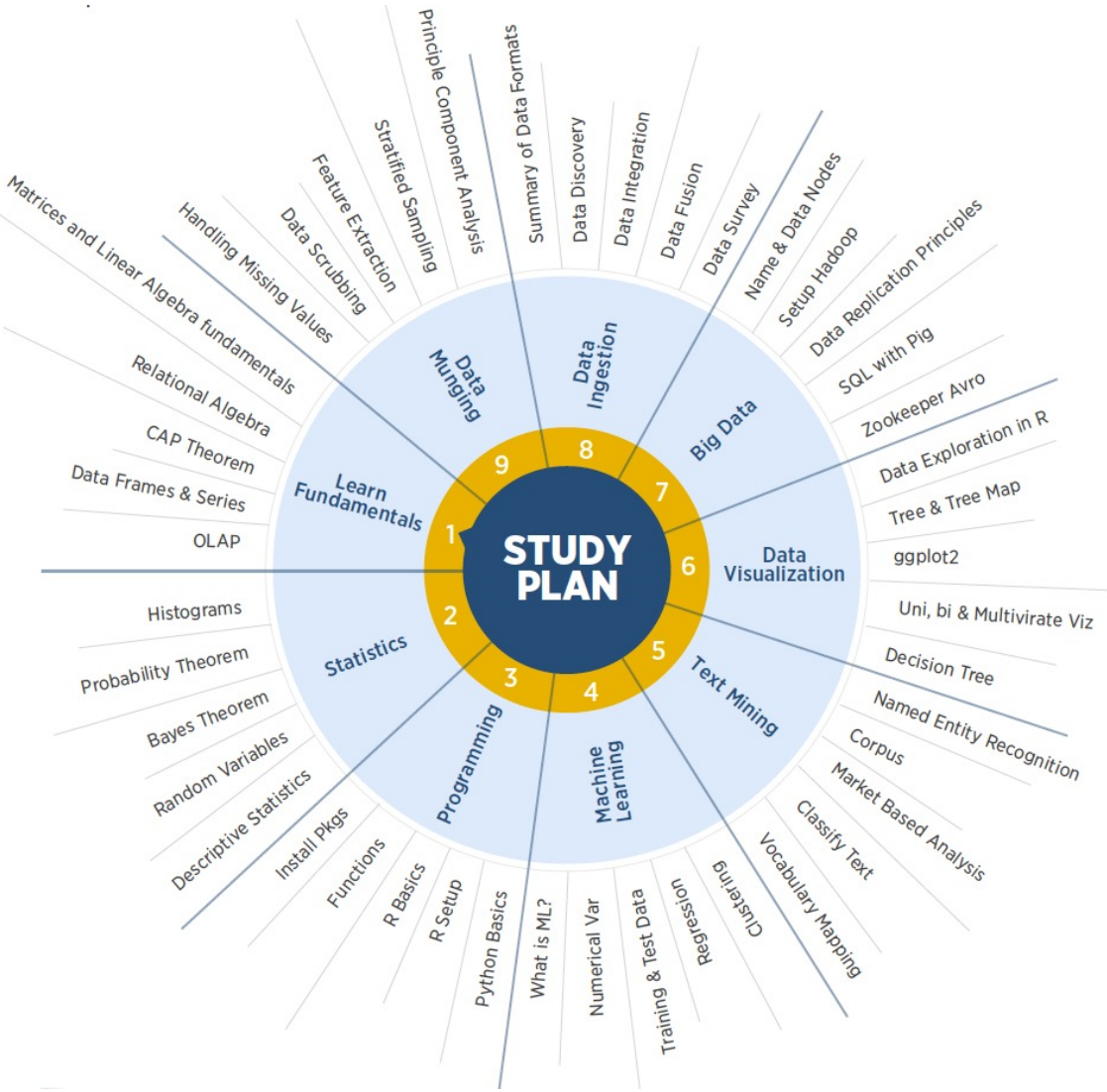
Προκαταρκτικά (1)

- **… λίγα λόγια για το τι είναι Data Science (και την ιστορία του όρου)**
  - As there is more and more data churned out every day, there is an increased need of procuring this data and making it useful. **Data Science** refers to **collection, preparation, analysis, visualization, management, and preservation of large collections of data**.
  - In simple terms, data science is the **extraction of useful information from the available data**. The methods that usually deal with Big Data are of particular interest in data science, though the latter is not restricted to such data.
  - In 1997, C.F Jeff Wu gave an inaugural lecture on "**Statistics = Data Science?**" at the Univ. Michigan. In this lecture, the term 'data science' was coined and it was advocated that statistics should be renamed data science and statisticians should be renamed data scientist.
  - In 2008, the term **Data Scientist** was coined by DJ Patil and Jeff Hammerbacher to define their jobs at LinkedIn and Facebook, respectively.

Προκαταρκτικά (2)

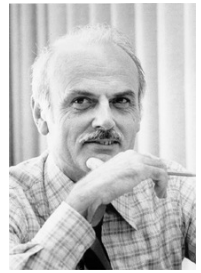
- **Τι πρέπει να γνωρίζει ένας Data Scientist**
  - Προγραμματισμό, Στατιστική, Διαχείριση (Μεγάλων) Δεδομένων, Οπτικοποίηση Δεδομένων, Μηχανική Μάθηση κλπ.

- *Data Science is expected to mature, consolidate, become the mainstream career option, and even surprise us with the advancements in the field. The field is expected to mature over time and slowly shift to cloud environment. Data science practitioners should be able to build predictive models in temporary cloud environments in order to increase their performance requirements. Unlike the current trend of single algorithm or tool being used to solve most of the data-related problems, the future is predicted to break this jinx.* ***Data scientists are building new data algorithms to suit their needs, which are predicted to take advantage of parallel data processing to improve efficiency.*** (Simplilearn.com, 2014)



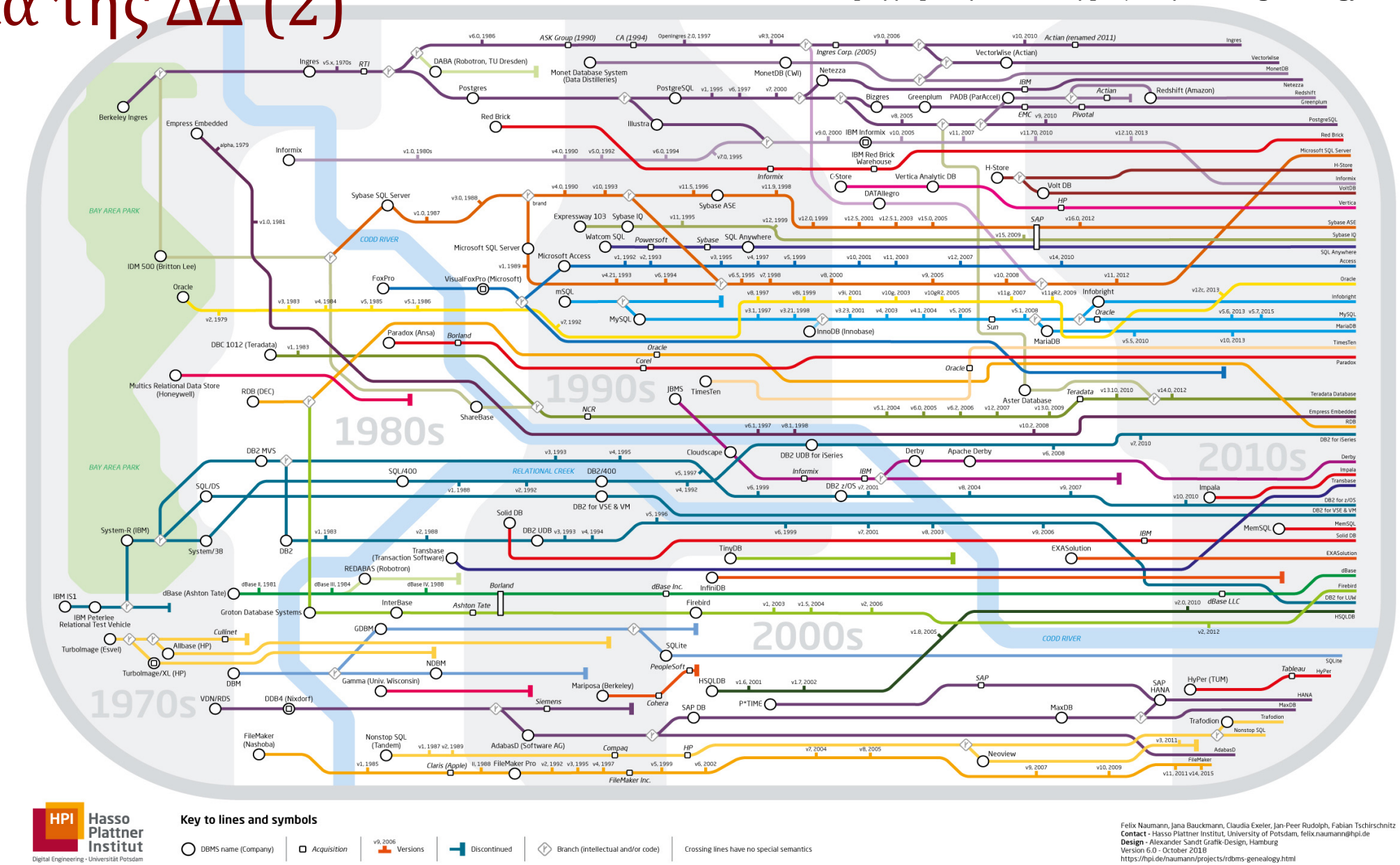
https://www.simplilearn.com/the-numbers-game-deciphered-guide-pdf

Η ιστορία της ΔΔ (1)



- **1970: Codd's paper -- the relational model (Turing award)**
  - **Data independence.** Allows for schema and physical storage structures to change under the covers
- **70's - 80's: From prototypes to commercial DBMS**
  - **Ingres** @ UC Berkeley → Sybase, MS SQL Server
  - **System R** @ IBM San Jose (now IBM Almaden) → IBM's DB2, Oracle
  - **SQL** becomes de-facto standard
- **90's: the age of maturity**
  - RDBMS improvements in terms of **transactional** facilities, **performance** and stability
  - **Objects**: object-relational modeling ➔ ORDBMS (support for ADT's)
  - **Parallel**, **Distributed** DBMS (scalability)
- **00' s - today: more scalability → Cloud, Big Data, IoT/edge**

Η ιστορία της ΔΔ (2)



Θέματα αιχμής στη ΔΔ (1)

- "… Senior database researchers gather every few years to **assess the state of database research** and to **recommend problems and problem areas that deserve additional focus**. …"
  - Laguna Beach, CA, 1989
  - Palo Alto, CA, 1990, 1995
  - Cambridge, MA, 1996
  - Asilomar, CA, 1998
  - Lowell, MA, 2003
  - Claremont, CA, 2008
  - Beckman, Irvine – CA, 2013
  - Seattle, WA, 2018

Group photo from Beckman meeting (2013)

Θέματα αιχμής στη ΔΔ (2)

- Asilomar report (1998): … broadening the definition of database management to embrace all the **content of the Web and other online data stores**, and rethinking our fundamental assumptions in light of technology shifts.
- Lowell report (2003): … **integration of text, data, code, and streams; fusion of information from heterogeneous data sources**; reasoning about uncertain data; unsupervised data mining for interesting correlations; **information privacy**; and self-adaptation and repair.
- Claremont report (2008): … new database engine architectures, declarative programming languages, the **interplay of structured and unstructured data, cloud data services, and mobile and virtual worlds**.
- Beckman report (2013): … the report recommends significantly more attention to five research areas: **scalable big/fast data infrastructures**; **coping with diversity in the data management landscape**; **end-to-end processing and understanding of data**; **cloud services**; and **managing the diverse roles of people on the data life-cycle**.
- Seattle report (2018): … Today, we are living in a data- driven society where decisions are increasingly driven by the insights gathered from analysis of relevant data ("**data is the new oil**"). … However, the fact that data is at the center of everything today also means that the field has grown in breadth and that new challenges have arisen.

Θέματα αιχμής στη ΔΔ (3)

Research challenges:

- **Data Science**: … and data scientists rely on a rich ecosystem of open-source libraries for sophisticated analysis, including the latest ML techniques
- **Data Governance**: Data-intensive applications that use [sensor, virtual assistant, social, etc.] data sources raise not only technical challenges but also those of privacy and ownership. Data producers have an economic and personal interest that the data is used only in certain ways.
- **Cloud Services**: Challenges of new consumption models (IaaS, etc.), challenges of cloud architecture (disaggregation, multi-tenancy, hybrid cloud, edge and cloud), auto tuning, confidential cloud computing, etc.
- **Database Engines**: We see a clear trend towards heterogeneous computation with the death of Dennard scaling and the advent of new accelerators introduced to offload compute. GPUs and FP- GAs are available today …

The Seattle Report on Database Research

Daniel Abadi, Anastasia Ailamaki, David Andersen, Peter Bailis, Magdalena Balazinska, Philip Bernstein, Peter Boncz, Surajit Chaudhuri, Alvin Cheung, AnHai Doan, Luna Dong, Michael J. Franklin, Juliana Freire, Alon Halevy, Joseph M. Hellerstein, Stratos Idreos, Donald Kossmann, Tim Kraska, Sailesh Krishnamurthy, Volker Markl, Sergey Melnik, Tova Milo, C. Mohan, Thomas Neumann, Beng Chin Ooi, Fatma Ozcan, Jignesh Patel, Andrew Pavlo, Raluca Popa, Raghu Ramakrishnan, Christopher Ré, Michael Stonebraker and Dan Suciu

ABSTRACT

Approximately every five years, a group of database researchers meet to do a self-assessment of our community, including reflections on our impact on the industry as well as challenges facing our research community. This report summarizes the discussion and conclusions of the 9th such meeting, held during October 9-10, 2018 in Seattle.

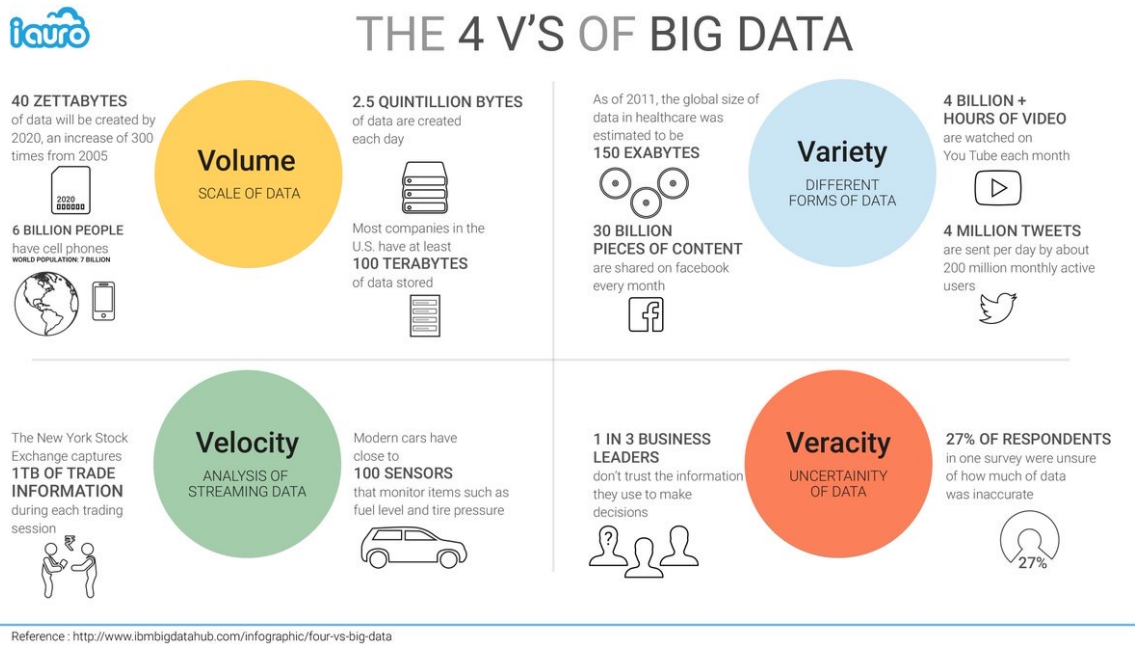
evolution of streaming data platforms as well as NoSQL systems.

Our achievements show that the state of our community is strong. Yet, in technology, the only constant is change. Today, we are living in a data-driven society where decisions are increasingly driven by the insights gathered from analysis of relevant data ("data is the new oil"). This societal trans-

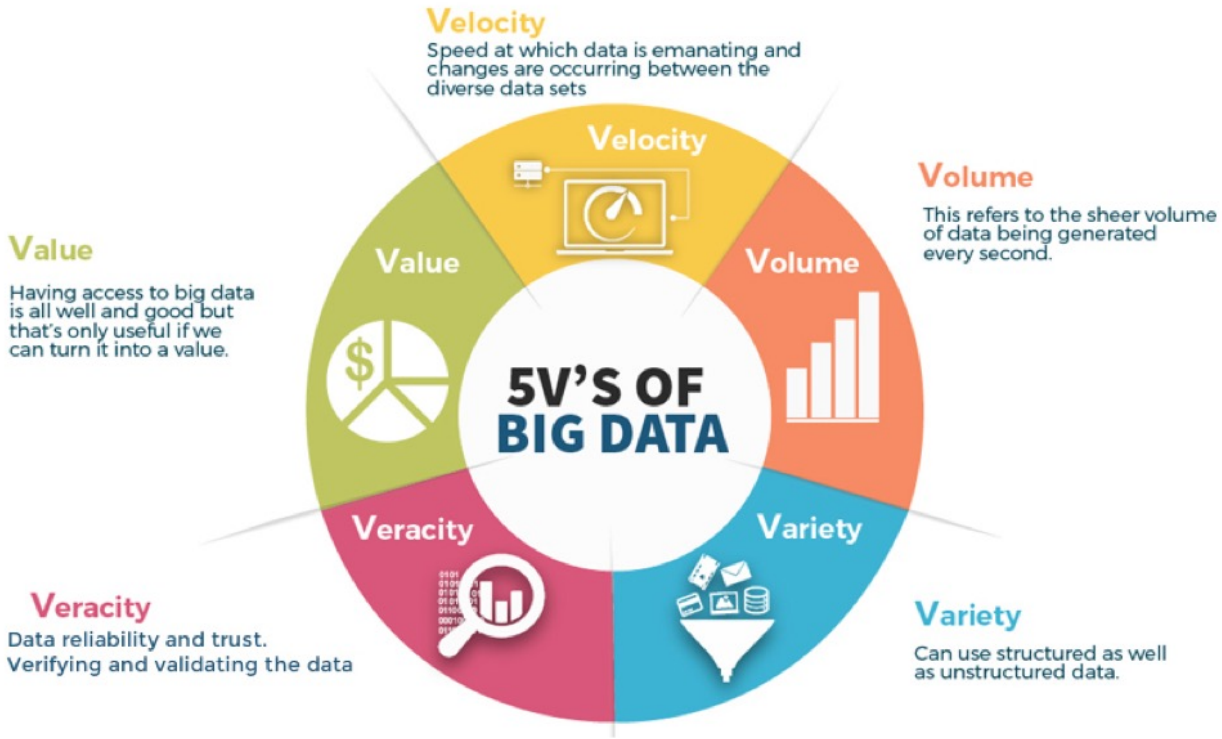
Looking forward: *Much has already changed since our Fall 2018 meeting. Every new mechanism that has emerged offers a potential opportunity to enhance data management capabilities (e.g., blockchain, quantum computing) and every new scenario is a potential application area where data management might help (e.g., self- driving cars, fake news) …*

Η εποχή των «Μεγάλων Δεδομένων» (1)

- **Τα 4 (5) V's**
  - Volume; Velocity; Variety; Veracity (; Value)



https://twitter.com/iauro

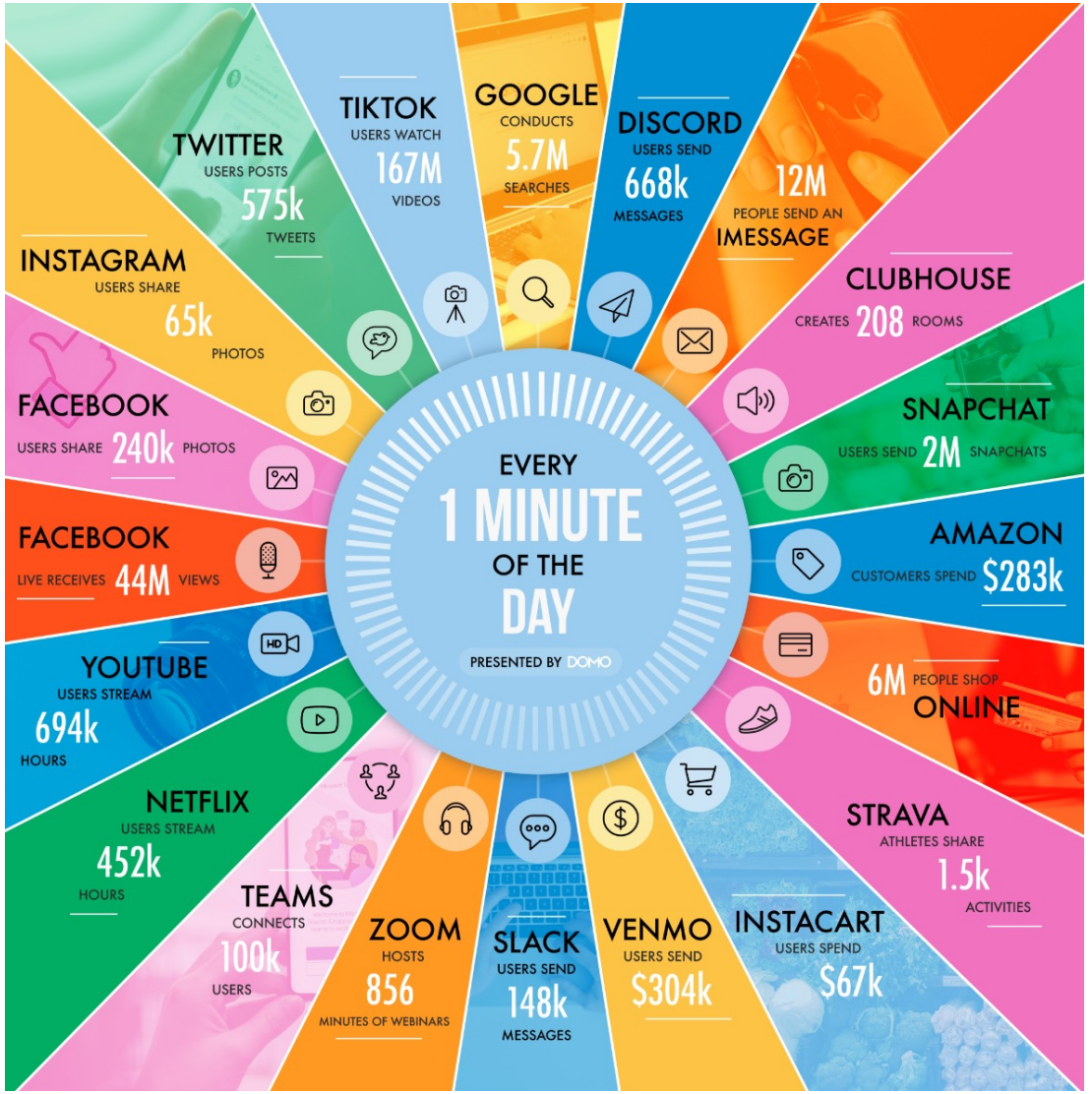


https://www.techentice.com/the-data-veracity-big-data/

Η εποχή των «Μεγάλων Δεδομένων» (2)

Every minute of the day (2021) …

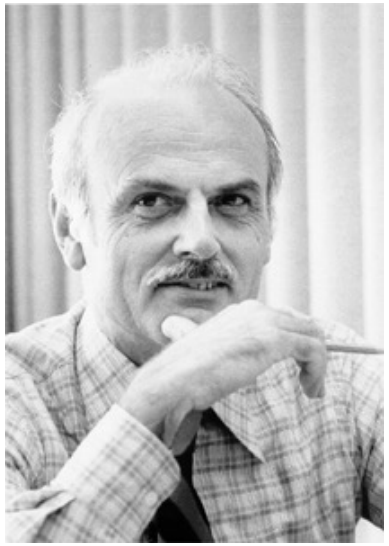
- **Facebook** users share 240K photos
- **Twitter** users post 575K tweets
- **Instagram** users share 65K photos
- **YouTube** users stream 694K hours
- **Amazon** customers spend $283K
- **Tiktok** users watch 167M videos
- **Netflix** users stream 452K hours
- etc.



https://www.domo.com/learn/infographic/data-never-sleeps-9

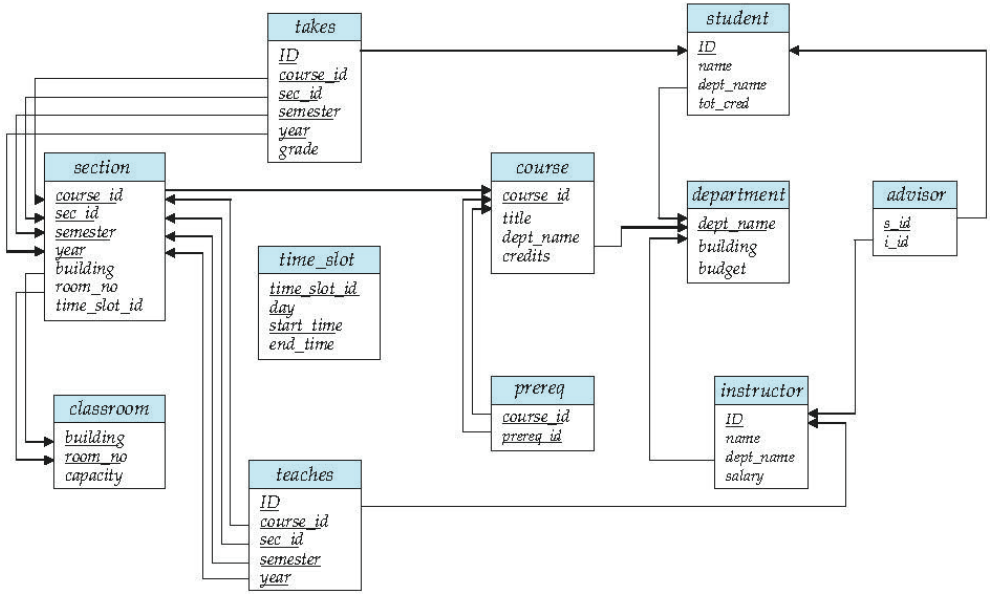
Μία γρήγορη επανάληψη στα «παραδοσιακά» (Σχεσιακά) Συστήματα Διαχείρισης Βάσεων Δεδομένων

E.F. “Ted” Codd (1923-2003)
Turing award 1981



Το Σχεσιακό Μοντέλο

- Η ΒΔ αποτελείται από ένα σύνολο **σχέσεων** (relations) ή **πινάκων** (tables) που συνδέονται κατάλληλα μεταξύ τους
- Σχέση: ένα σύνολο **πλειάδων** (εγγραφών, γραμμών) ορισμένων βάσει κάποιων **χαρακτηριστικών** (πεδίων, στηλών)



** Codd, E. F. (1970). "A relational model of data for large shared data banks". Communications of the ACM 13 (6): 377-387.*

A Relational Model of Data for Large Shared Data Banks

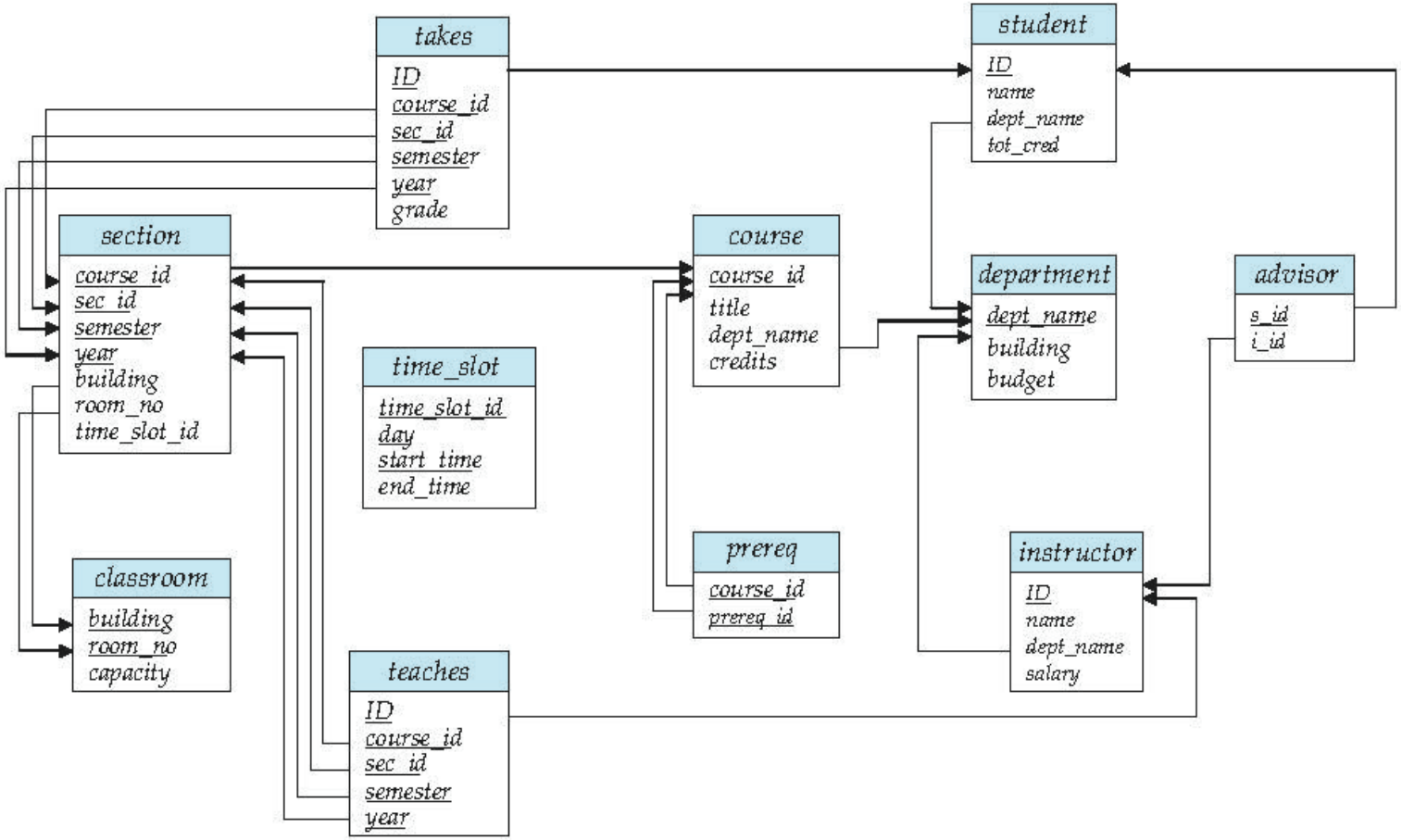
E. F. Codd
IBM Research Laboratory, San Jose, California

Future users of large data banks must be protected from having to know how the data is organized in the machine (the internal representation). A prompting service which supplies such information is not a satisfactory solution. Activities of users at terminals and most application programs should remain unaffected when the internal representation of data is changed and even when some aspects of the external representation are changed. Changes in data representation will often be

The relational view (or model) of data described in Section 1 appears to be superior in several respects to the graph or network model [3, 4] presently in vogue for non-inferential systems. It provides a means of describing data with its natural structure only—that is, without superimposing any additional structure for machine representation purposes. Accordingly, it provides a basis for a high level data language which will yield maximal independence between programs on the one hand and machine representation and organization of data on the other.

A further advantage of the relational view is that it forms a sound basis for treating derivability, redundancy, and consistency of relations—these are discussed in Section 2. The network model, on the other hand, has spawned a number of confusions, not the least of which is mistaking the derivation of connections for the derivation of relations (see remarks in Section 2 on the "connection trap").

Παράδειγμα σχεσιακής ΒΔ



ID	name	dept_name	salary
22222	Einstein	Physics	95000
12121	Wu	Finance	90000
32343	El Said	History	60000
45565	Katz	Comp. Sci.	75000
98345	Kim	Elec. Eng.	80000
76766	Crick	Biology	72000
10101	Srinivasan	Comp. Sci.	65000
58583	Califieri	History	62000
83821	Brandt	Comp. Sci.	92000
15151	Mozart	Music	40000
33456	Gold	Physics	87000
76543	Singh	Finance	80000

(a) The *instructor* table

dept_name	building	budget
Comp. Sci.	Taylor	100000
Biology	Watson	90000
Elec. Eng.	Taylor	85000
Music	Packard	80000
Finance	Painter	120000
History	Painter	50000
Physics	Watson	70000

(b) The *department* table

ακεραιότητα οντότητας (entity integrity) ➔ η έννοια του **πρωτεύοντος κλειδιού** (primary key)

αναφορική ακεραιότητα (referential integrity) ➔ η έννοια του **ξένου κλειδιού** (foreign key)

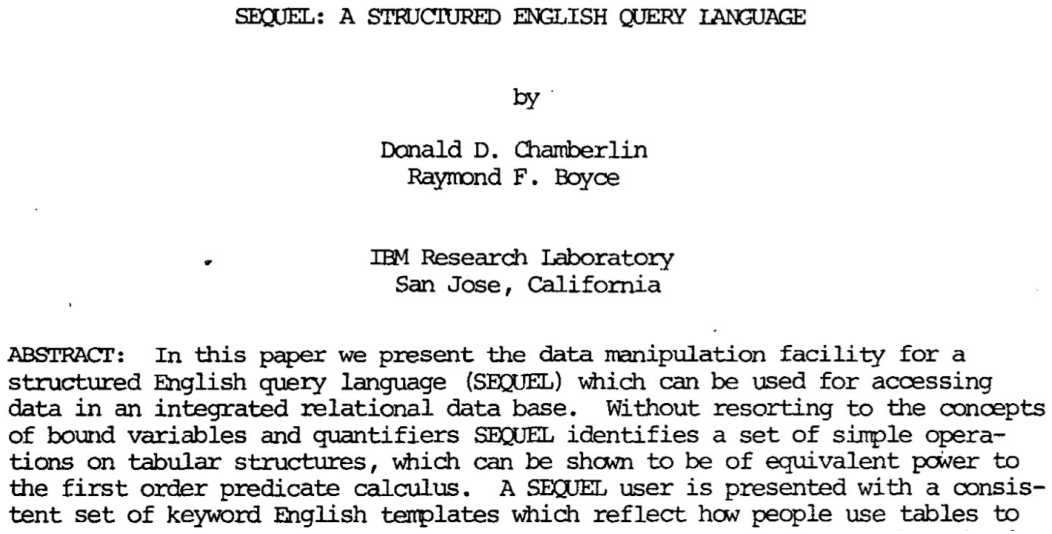
Η γλώσσα SQL

- Η SQL βασίζεται σε πράξεις επιλογής-προβολής-σύνδεσης (select-project-join -- SPJ) μεταξύ σχέσεων
  - θεωρητικό υπόβαθρο: Σχεσιακή Άλγεβρα (Relational Algebra)

```
SELECT  DISTINCT I.name, C.title
 FROM   instructor I
        INNER JOIN teaches T USING ID
        INNER JOIN course C USING course_id
 WHERE C.dept_name = "Music" AND T.year = 2009
```

$$\Pi_{name,title}(\sigma_{dept\_name = \text{“Music”} \text{ AND } year = 2009} ((instructor \bowtie teaches) \bowtie course))$$

** Chamberlin, Donald D; Boyce, Raymond F (1974). "SEQUEL: A Structured English Query Language". Proceedings of the 1974 ACM SIGFIDET Workshop on Data Description, Access and Control: 249–64.*



SEQUEL: A STRUCTURED ENGLISH QUERY LANGUAGE

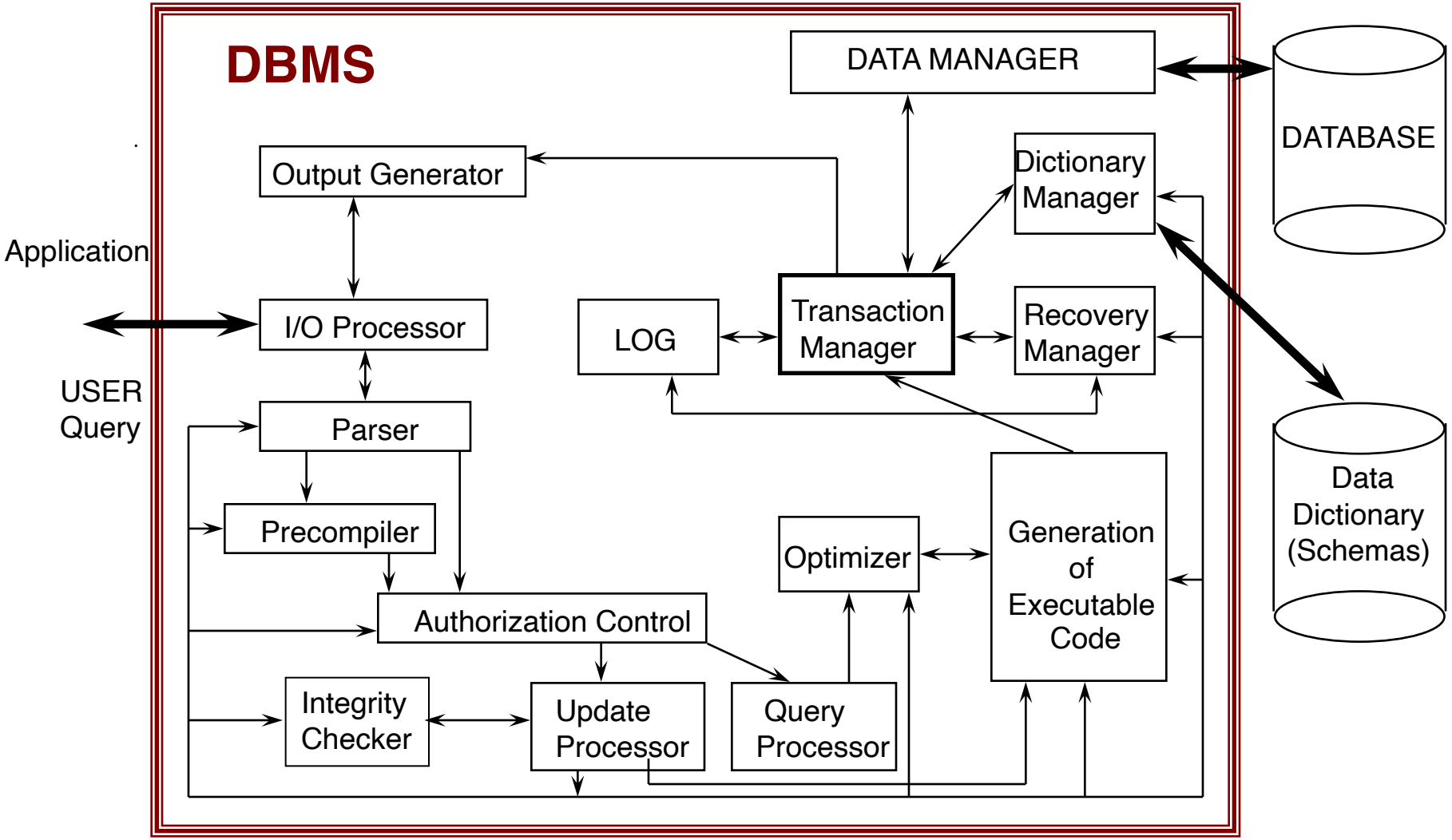
by

Donald D. Chamberlin
Raymond F. Boyce

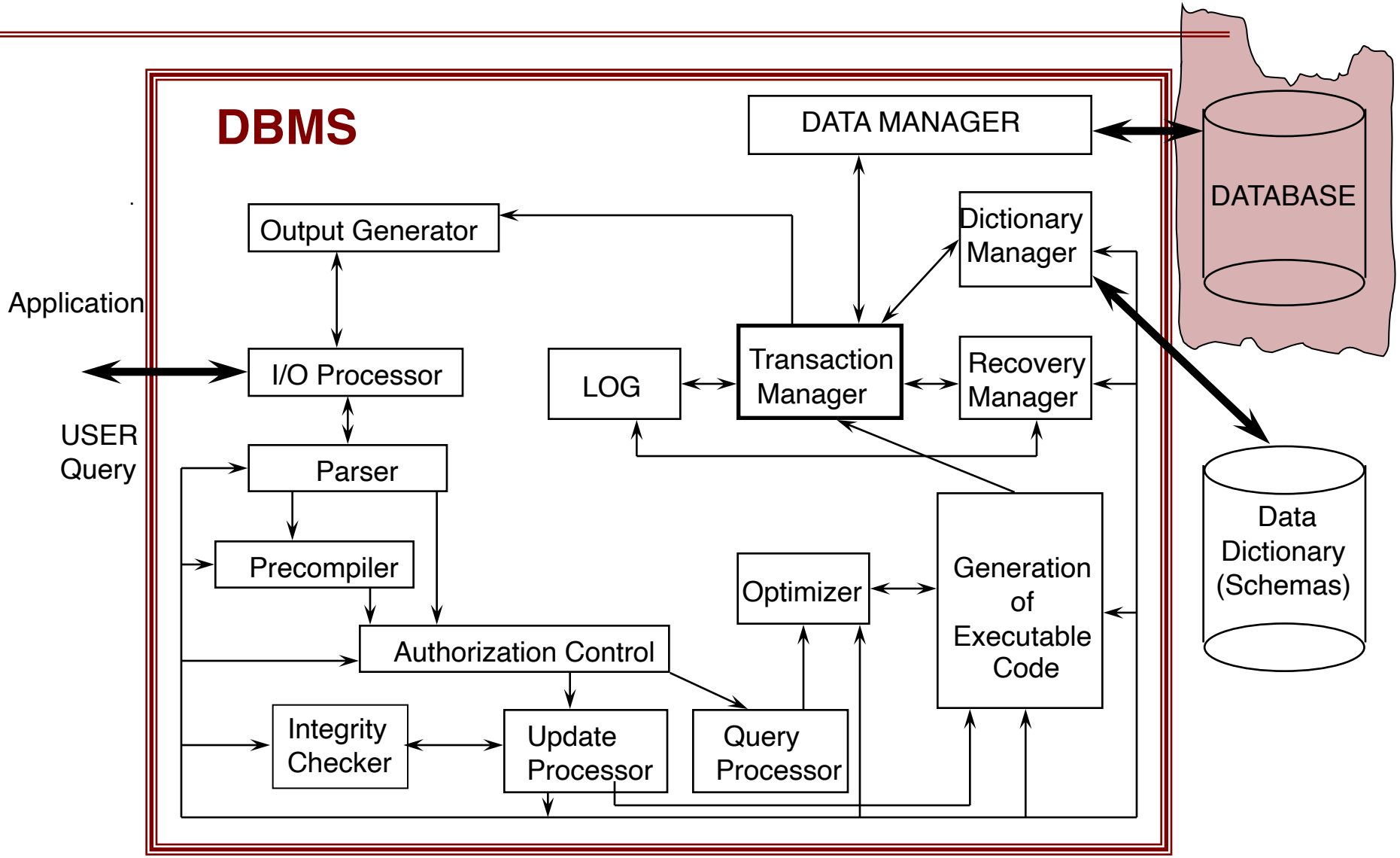
IBM Research Laboratory
San Jose, California

ABSTRACT: In this paper we present the data manipulation facility for a structured English query language (SEQUEL) which can be used for accessing data in an integrated relational data base. Without resorting to the concepts of bound variables and quantifiers SEQUEL identifies a set of simple operations on tabular structures, which can be shown to be of equivalent power to the first order predicate calculus. A SEQUEL user is presented with a consistent set of keyword English templates which reflect how people use tables to

Αρχιτεκτονική ενός ΣΔΒΔ



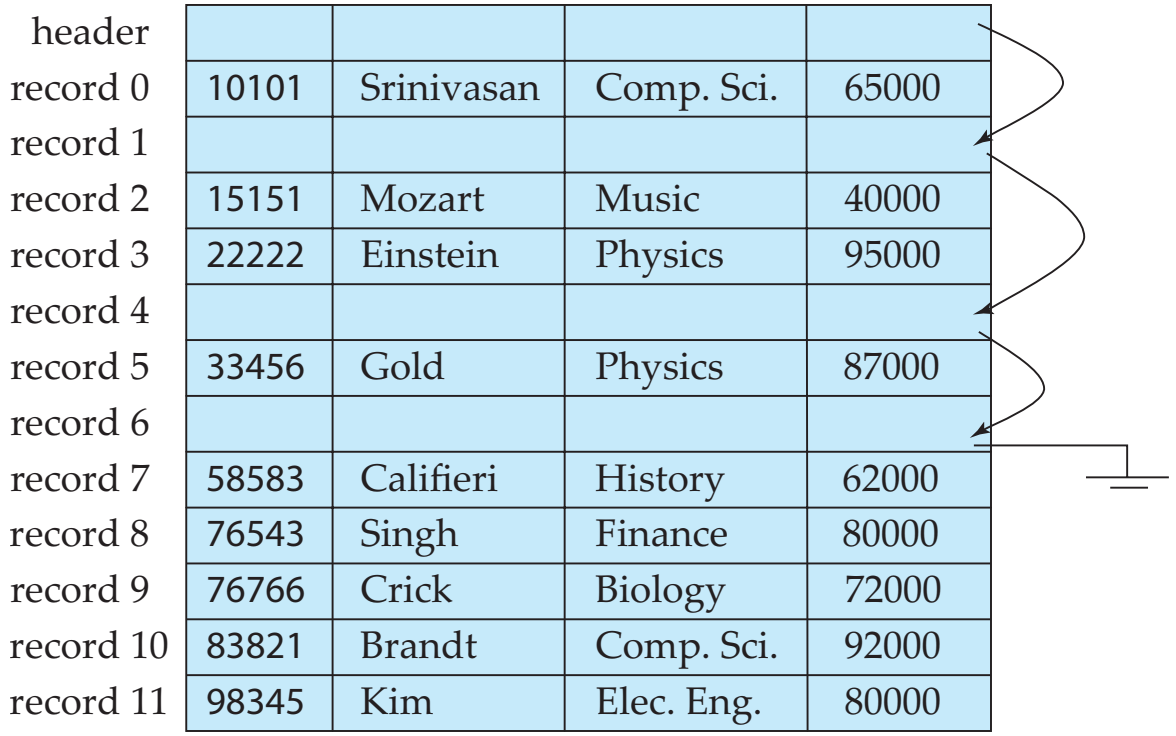
Αποθήκευση στο φυσικό μέσο ...



Οργάνωση αρχείων & Ευρετήρια (1)

- Αρχείο σωρού vs.
  Ακολουθιακό (Διατεταγμένο) vs.
  Κατακερματισμένο
  - Εσωτερικά οι εγγραφές έχουν κάποια **διάταξη** ή όχι;
  - Σε τι εξυπηρετούν οι ειδικές οργανώσεις;

- Εκτός από το κύριο αρχείο, υπάρχουν άλλες βοηθητικές δομές;
  - **Ευρετήρια** (B+δέντρα κ.α.)

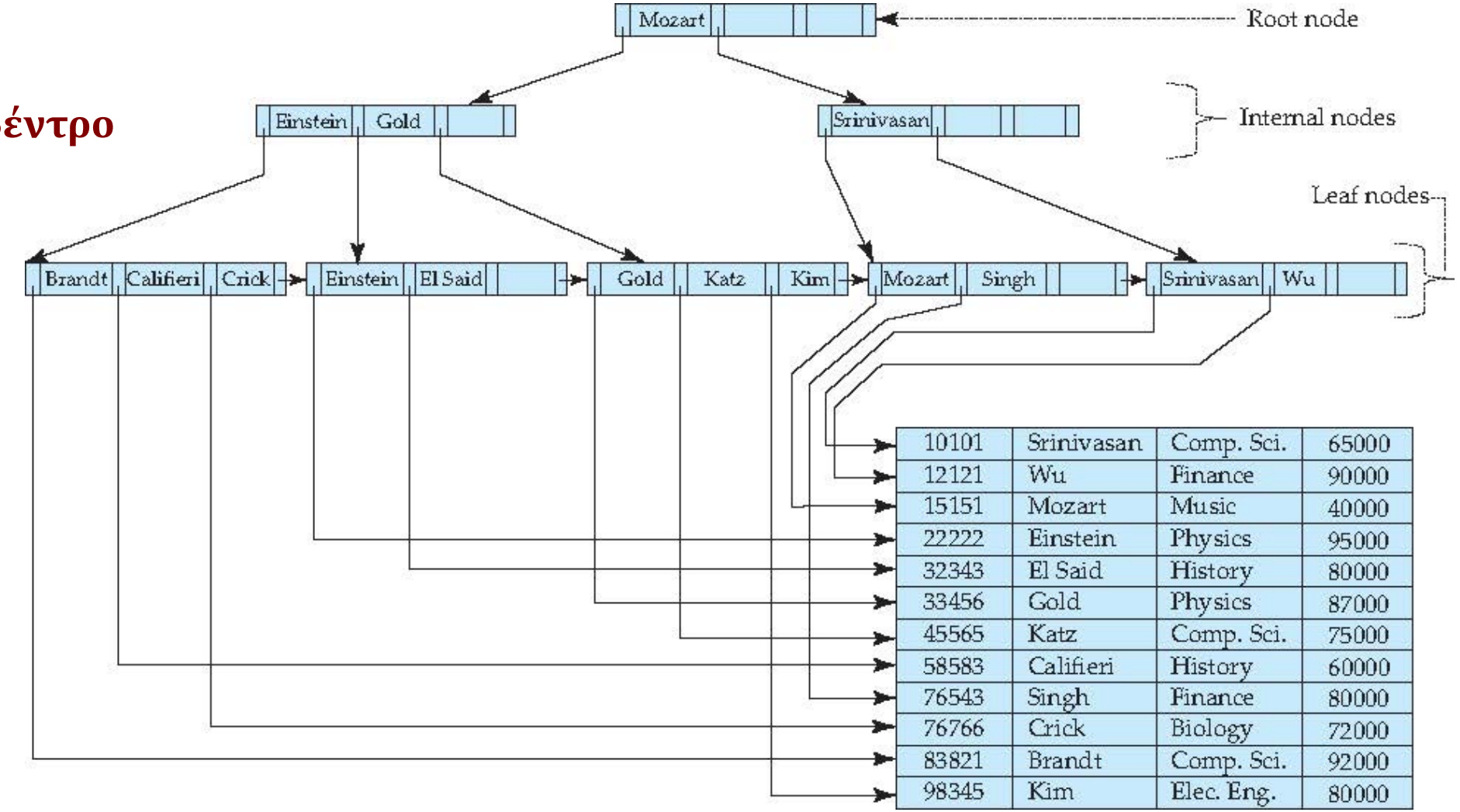


header				
record 0	10101	Srinivasan	Comp. Sci.	65000
record 1				
record 2	15151	Mozart	Music	40000
record 3	22222	Einstein	Physics	95000
record 4				
record 5	33456	Gold	Physics	87000
record 6				
record 7	58583	Califieri	History	62000
record 8	76543	Singh	Finance	80000
record 9	76766	Crick	Biology	72000
record 10	83821	Brandt	Comp. Sci.	92000
record 11	98345	Kim	Elec. Eng.	80000

Οργάνωση αρχείων & Ευρετήρια (2)

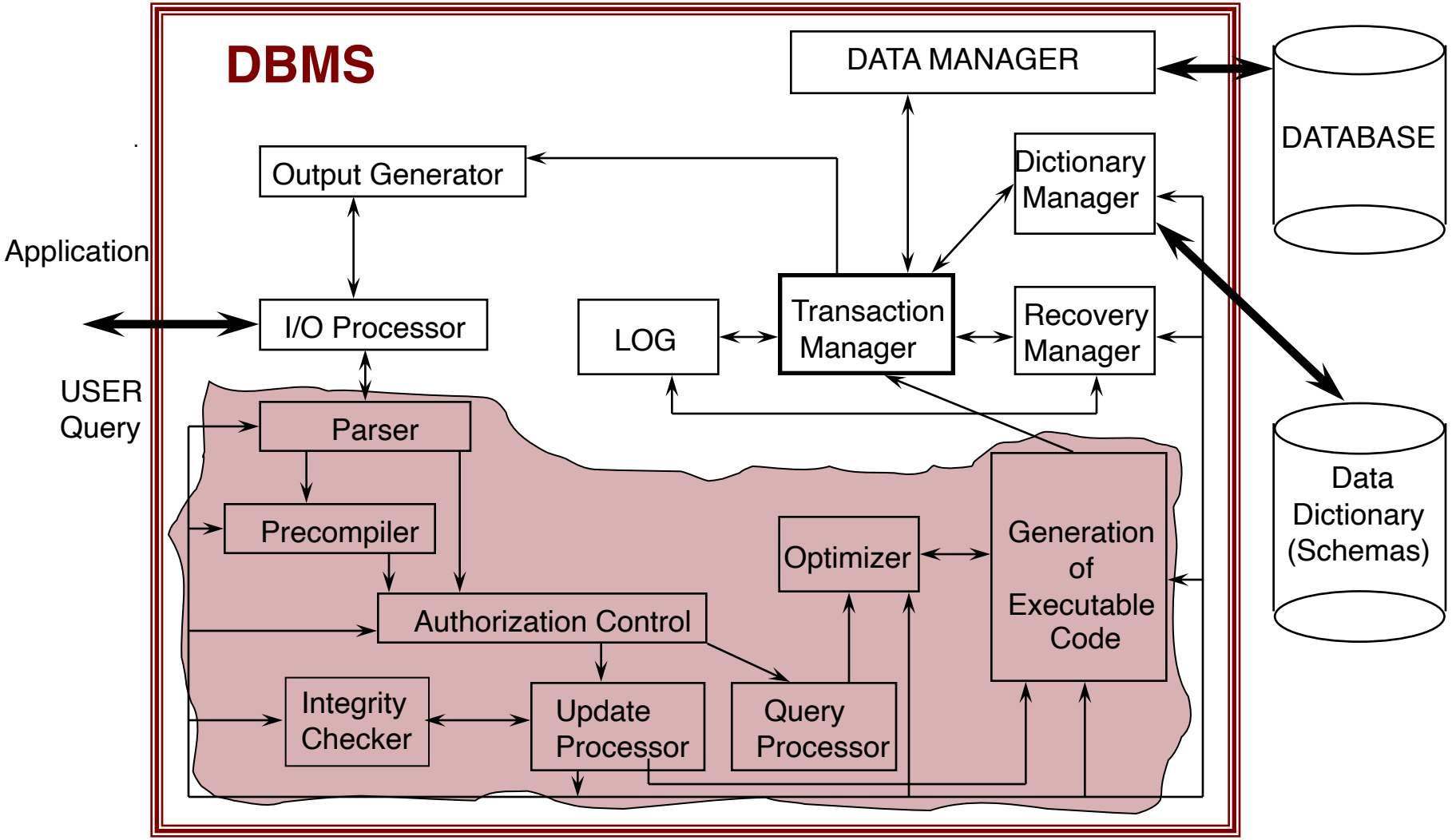
- **Ευρετήριο B+δέντρο**

Donald Knuth
(Stanford Univ.)



10101	Srinivasan	Comp. Sci.	65000
12121	Wu	Finance	90000
15151	Mozart	Music	40000
22222	Einstein	Physics	95000
32343	El Said	History	80000
33456	Gold	Physics	87000
45565	Katz	Comp. Sci.	75000
58583	Califieri	History	60000
76543	Singh	Finance	80000
76766	Crick	Biology	72000
83821	Brandt	Comp. Sci.	92000
98345	Kim	Elec. Eng.	80000

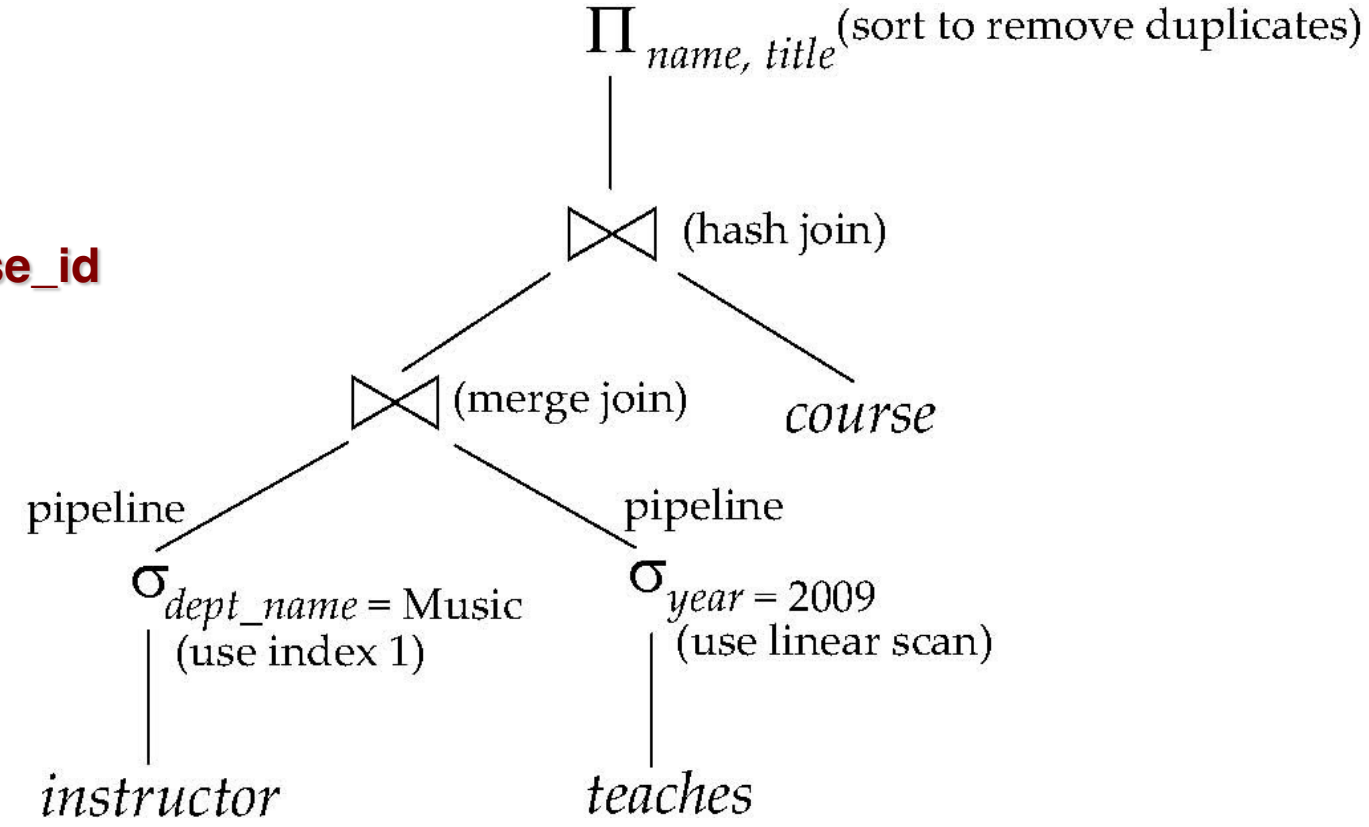
Αποδοτική εκτέλεση εντολών ...



Βελτιστοποιητής ερωτήσεων (Query optimizer)

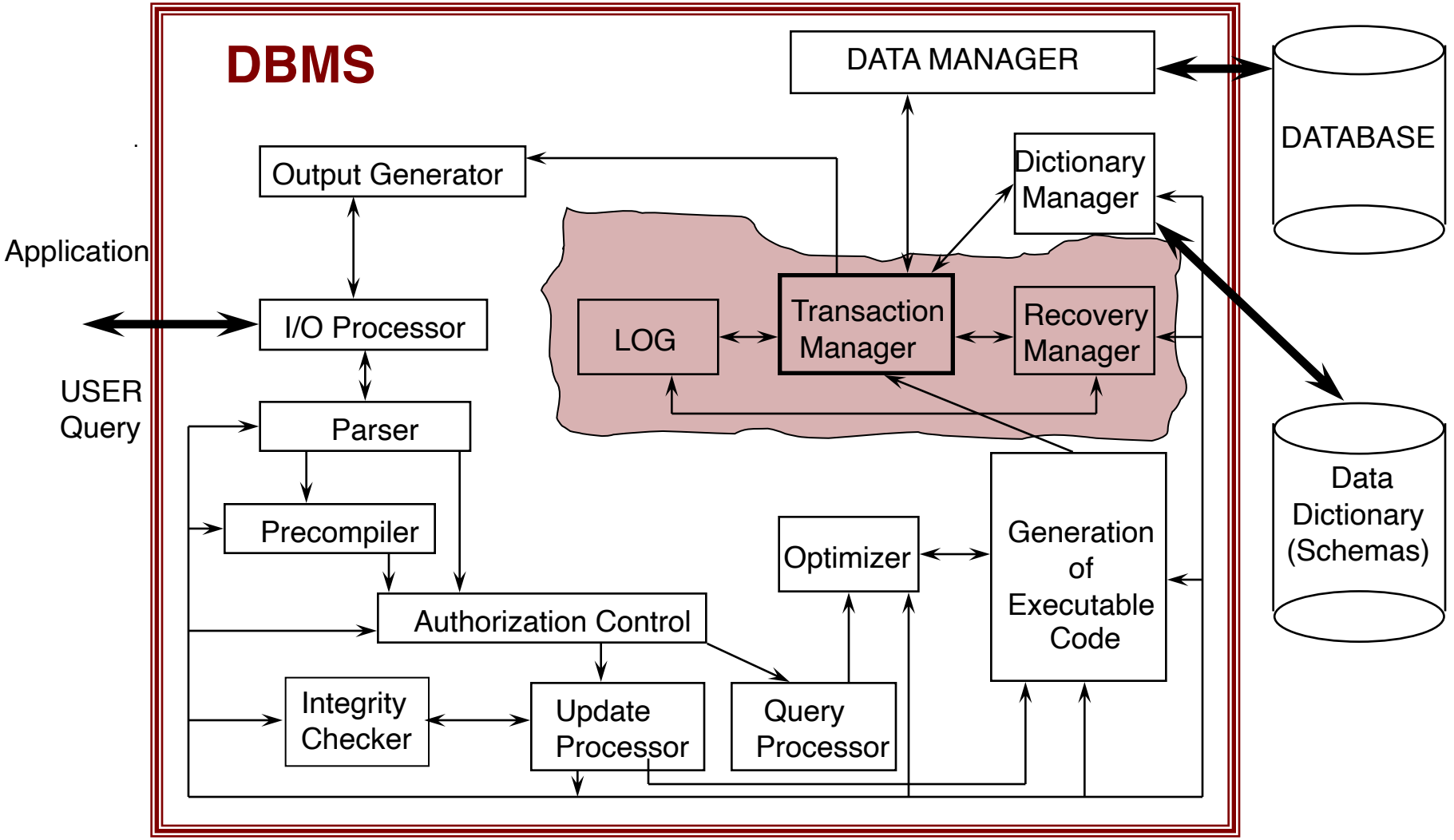
- Όταν ο χρήστης υποβάλει μια ερώτηση στο DBMS στη (δηλωτική) γλώσσα SQL, το DBMS αποφασίζει να εκτελέσει ένα **πλάνο εκτέλεσης** (query plan)
- Παράδειγμα πλάνου εκτέλεσης:

```
SELECT I.name, C.title
  FROM instructor I
  INNER JOIN teaches T USING ID
  INNER JOIN course C USING course_id
  WHERE C.dept_name = “Music”
    AND T.year = 2009
```



- Γιατί αυτό και όχι κάποιο άλλο πλάνο εκτέλεσης;

Εξασφάλιση συνέπειας ΒΔ ...



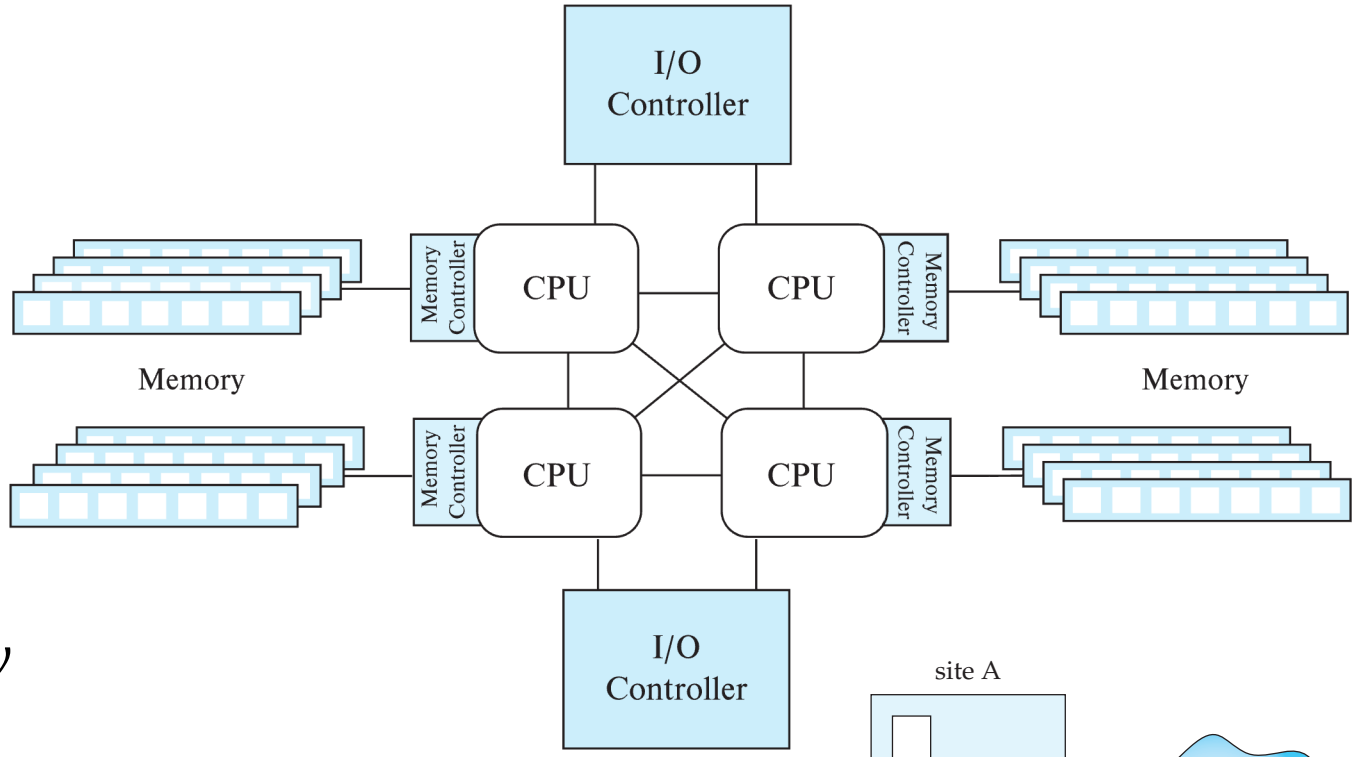
Διαχείριση δοσοληψιών

- Μια **δοσοληψία** ή **συναλλαγή** (transaction) είναι ένα τμήμα της εκτέλεσης προγράμματος, που διαβάζει και πιθανόν ενημερώνει δεδομένα της ΒΔ.
- Παράδειγμα: μεταφορά πιστωτικών μονάδων μεταξύ δύο φοιτητών (λόγω π.χ. εκ παραδρομής βαθμολογίας)
  - **UPDATE Student SET tot-cred = tot-cred – 4 WHERE ID = "P12345"**
  - **UPDATE Student SET tot-cred = tot-cred + 4 WHERE ID = "P12354"**
- Πρέπει να διατηρηθεί η «συνέπεια» της ΒΔ ανεξαρτήτως των όποιων προβλημάτων προκύψουν
  - πτώση συστήματος, ταυτόχρονη εκτέλεση πολλαπλών δοσοληψιών κ.α.

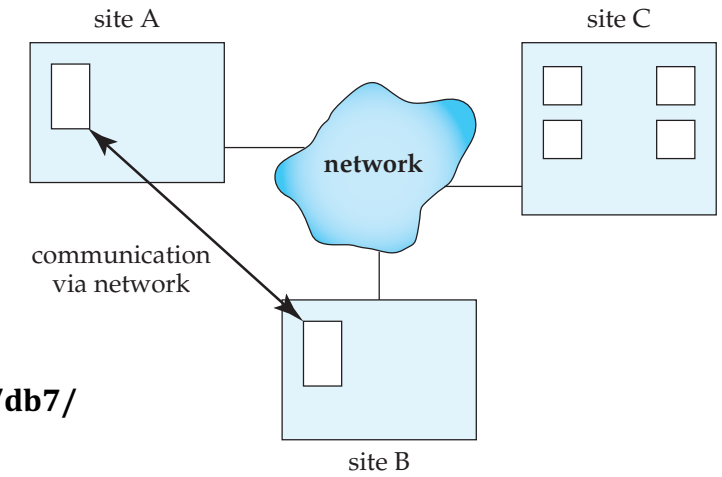
1. **read**(*A*)
2. *A* := *A* – 4
3. **write**(*A*)
4. **read**(*B*)
5. *B* := *B* + 4
6. **write**(*B*)

Κατανεμημένες / Παράλληλες αρχιτεκτονικές

- Τα δεδομένα είναι κατανεμημένα σε πολλές μηχανές, συνδεδεμένες μέσω δικτύου
- Τα δεδομένα είναι κοινά για χρήστες που τα προσπελάζουν από διαφορετικές μηχανές
- Με κατάλληλους μηχανισμούς, επιτυγχάνεται παραλληλία στην επεξεργασία των δεδομένων



- Παράλληλες ΒΔ, Κατανεμημένες ΒΔ, ΒΔ στο υπολογιστικό νέφος (cloud), κλπ.



https://www.db-book.com/db7/

Περαιτέρω μελέτη

- Codd, E. F. (1970) **A Relational Model of Data for Large Shared Data Banks**. Comm. ACM 13(6): 377-387. URL: http://doi.acm.org/10.1145/362384.362685.
  - το κλασικό άρθρο απ' όπου ξεκίνησαν όλα !
- Stonebraker, M. (2015) **Traditional RDBMS Systems**. Chapter 2 in Readings in Database Systems, 5/e. URL: http://www.redbook.io.
  - μία επισκόπηση ιστορικά σημαντικών Σχεσιακών ΣΔΒΔ (System R, Postgres, Gamma)
- Ό,τι πουν οι «σοφοί» ...
  - D. Abadi et al. (2018) **The Seattle report on database research**. URL: https://sigmodrecord.org/publications/sigmodRecord/1912/pdfs/07_Reports_Abadi.pdf

- **Σημειώσεις, διαφάνειες, σχετικά άρθρα**: https://gunet2.cs.unipi.gr/courses/CDS110/
- **Ερευνητικά ενδιαφέροντα, ιδέες για διπλωματικές κλπ.** : https://www.datastories.org

Ευχαριστώ για την προσοχή σας!