Logistics

Instructor:

- Prof. Bertram Ludaescher (ludaesch@ucdavis.edu)
  - Office Hours: M, 12-1:30pm, 3051 Kemper Hall

Teaching Assistants:

- Megan Raul (meghanraul@me.com)
  - Office Hours: TBD 53/55 Kemper Hall
- Harini Sabbela (hhsabbella@ucdavis.edu)
  - Office Hours: TBD 53/55 Kemper Hall
- Steven Crites (sbcrites@ucdavis.edu)
  - Office Hours: TBD 53/55 Kemper Hall

sites.google.com/site/ecs165afq13

ECS-165A Database Systems

Search this site

- Home
- Lecture Notes
- Discussion Sections
- Assignments
- Grading, Policies
- Exams
- Mailing-List
- Resources

Home

Overview of topics:
Database design, Entity-Relationship model, Relational Model and Algebra, query language SQL, indexing, query processing, transaction management, hands-on exercises/application programming.

Instructor:
* Prof. Bertram Ludaescher (ludaesch@ucdavis.edu)
  Office Hours: M 12-1:30pm, 3051 Kemper Hall

Teaching Assistants:
* Megan Raul (meghanraul@me.com)
  Office Hours: TBD 53/55 Kemper Hall
* Harini Sabbela (hhsabbella@ucdavis.edu)
  Office Hours: TBD 53/55 Kemper Hall
* Steven Crites (sbcrites@ucdavis.edu)
  Office Hours: TBD

MEETING	CRN#	TIME	ROOM	Staff
Lecture		TR 12:10-1:30pm	2205 Haring	B. Ludäscher
Discussion Section	30768 30768	M 3:10-4pm F 2:10-3:00pm	206 Olson 115 Hutchison	TAs

Class Mailing List:
Sign-up: http://groups.google.com/group/165a-fq13
Email-to: 165a-fq13@googlegroups.com

Textbook
*** Database Systems: The Complete Book (2nd Edition),** Garcia-Molina, Ullman, Widom, Prentice Hall; 2nd ed. (2008)

More Logistics

Class page:

- sites.google.com/site/ecs165afq13

Mailing list:

- Sign up: groups.google.com/group/ecs165a-fq13
- Email: ecs165a-fq13@googlegroups.com

Textbook:

- ***Database Systems: The Complete Book (2nd Edition)*** by Garcia-Molina, Ullman, and Widom, Pearson Prentice-Hall, 2008/2009.

165A Course Topics

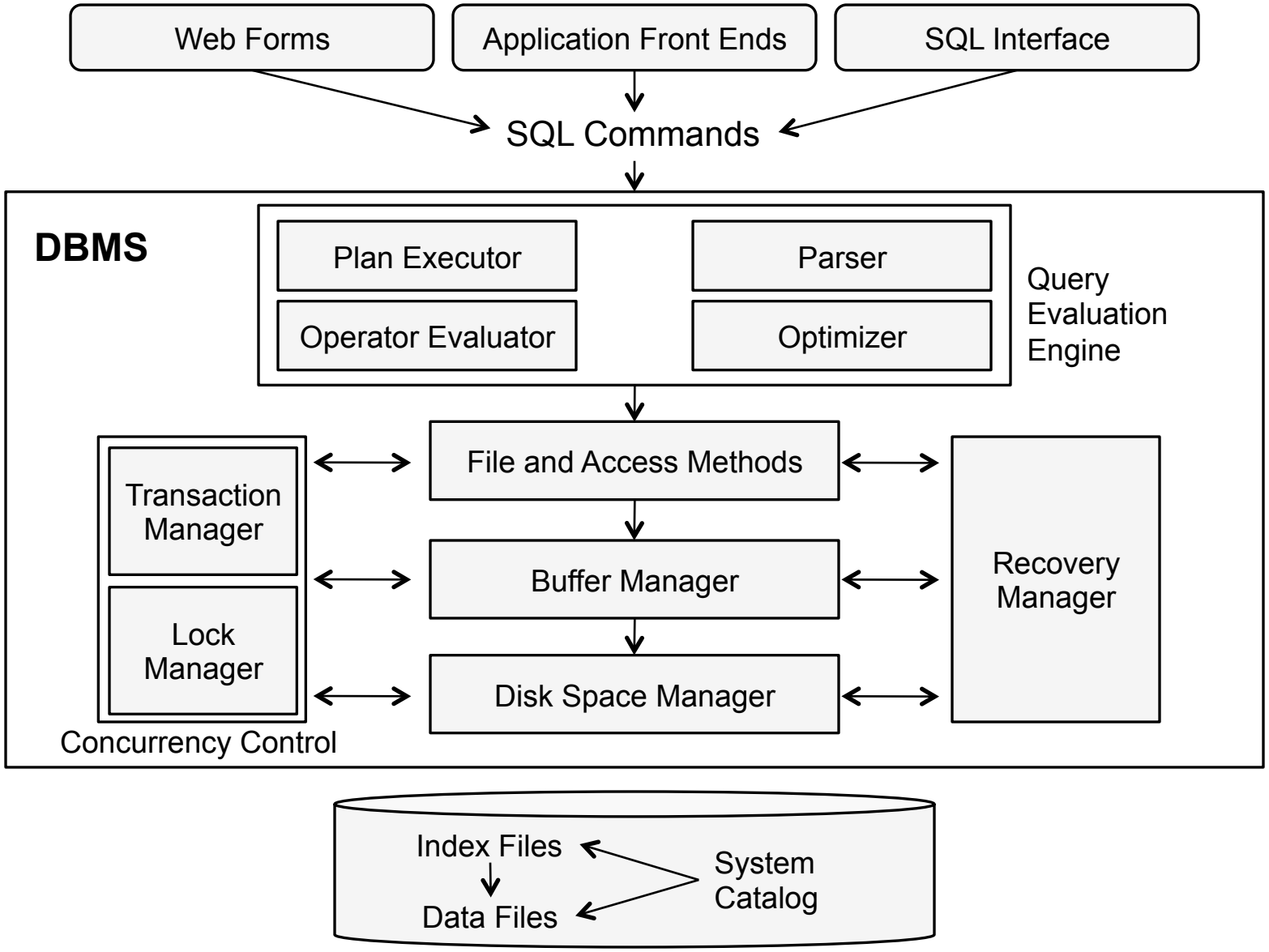
- Database Design, ER Model
- Relational Model, Relational Algebra
- SQL (Structured Query Language)
- Integrity Constraints
- Storage structures, Indexing
- Query Processing
- Transactions
- Additional Topics & Current Trends

165A Course Topics

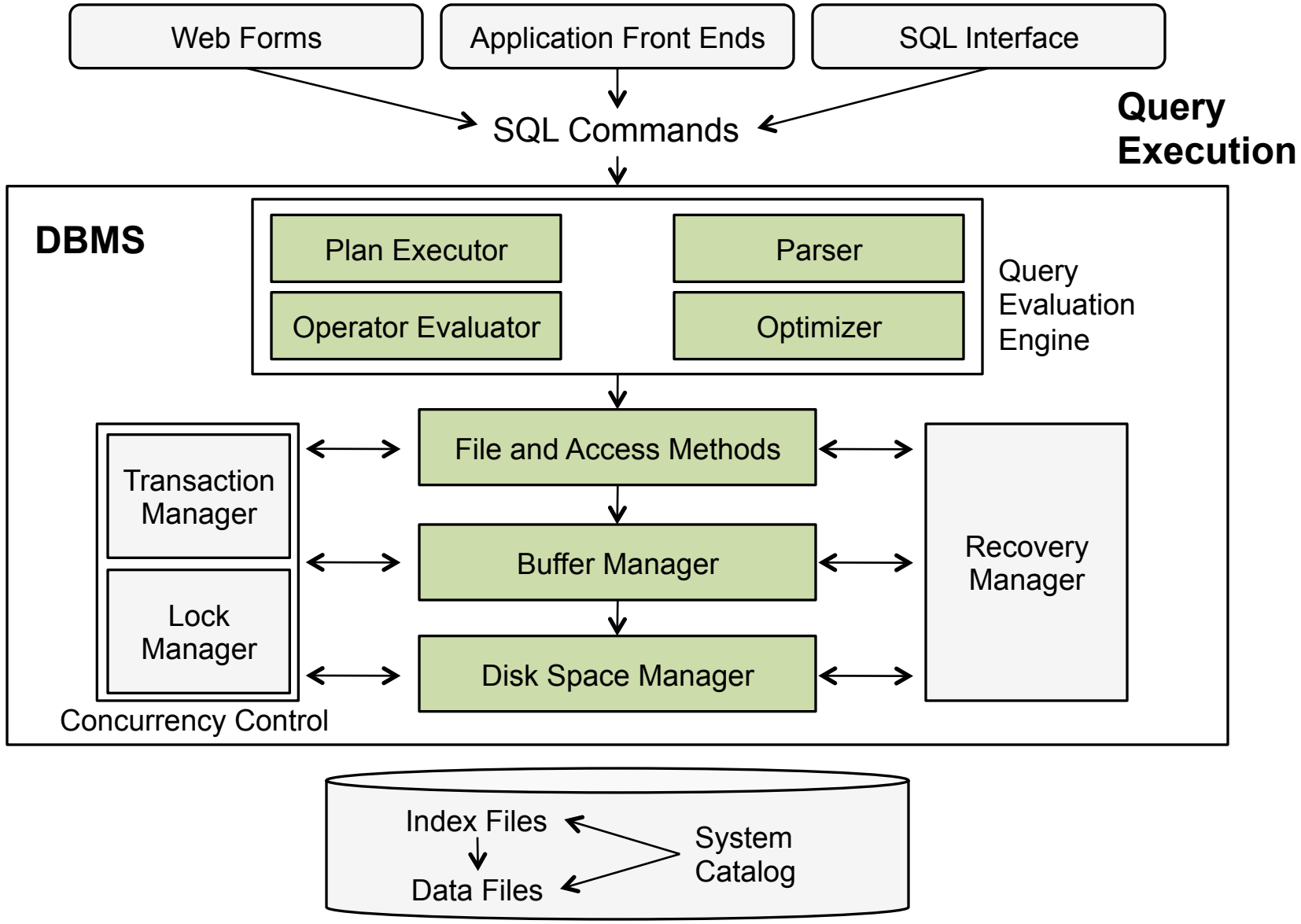
- Focus is on
  - **Foundations** (relational model, queries, SQL,…)
  - **Practical experience** with SQL
    - We'll use PostgreSQL
      - A "real" (full-featured), scablable DBMS
      - Open source, available @CSIF and @home!
        - » also looking at MySQL, SQLite, and
        - » Embedded SQL (e.g. with Python)

- **Individual** Assignments
- Group Project at the end

Basic Database Architecture



Query Processing



Grading and Policies

- Grading:
  - Approximately (see web page for details):
    - 40% Homework Assignments
    - 20% Midterm (also individual ;-)
    - 40% Final (and yes: this one too!
- Academic Conduct
  - Be polite
  - Don't cheat
- Ask when in doubt
- Make good use of the mailing-list!

Why study databases / data management?

- Critical to business, government, science, culture, society, …
- Determines success of many corporations (even their existence)
- Many tech companies built on data management (Google, Amazon, Yahoo!, Facebook, …)
- … or offer database products (Microsoft, IBM, Oracle)
- Database systems span major areas of computer science
  - Operating systems (file, memory, process management)
  - Theory (languages, algorithms, complexity)
  - Artificial Intelligence (knowledge-based systems, logic, search)
  - Software Engineering (application development)
  - Data structures (trees, hash-tables)
  - ***… and the DB research community continues to be very active***

Lots of Data Everywhere

- From http://en.wikipedia.org/wiki/Petabyte :

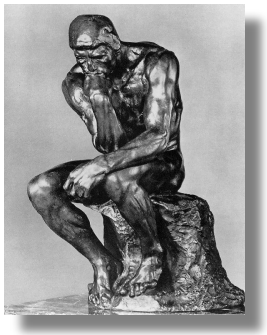
- **History**: According to Kevin Kelly in *The New York Times*, "the entire [written] works of humankind, from the beginning of recorded history, in all languages" would amount to 50 petabytes of data.[1]
- **Computer hardware**: Teradata Database 12 has a capacity of 50 petabytes of compressed data.[2][3]
- **Telecoms**: AT&T has about 16 petabytes of data transferred through their networks each day.[4]
- **Archives**: The Internet Archive contains about 3 petabytes of data, and is growing at the rate of about 100 terabytes per month as of March, 2009.[5][6]
- **Internet**: Google processes about 20 petabytes of data per day.[7]
- **Physics**: The 4 experiments in the Large Hadron Collider will produce about 15 petabytes of data per year, which will be distributed over the LHC Computing Grid.[8]
- **P2P networks**: As of October 2009, Isohunt has about 9.76 petabytes of files contained in torrents indexed globally.[9]
- **Games**: World of Warcraft utilizes 1.3 petabytes of storage to maintain its game.[10]

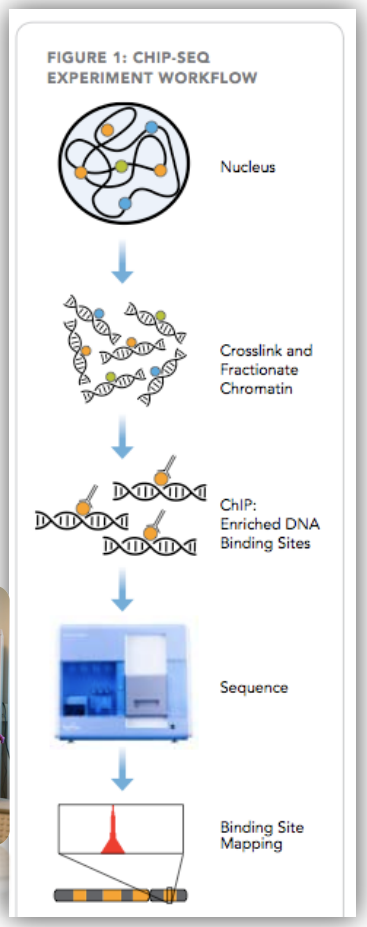
Science has been changing lately …

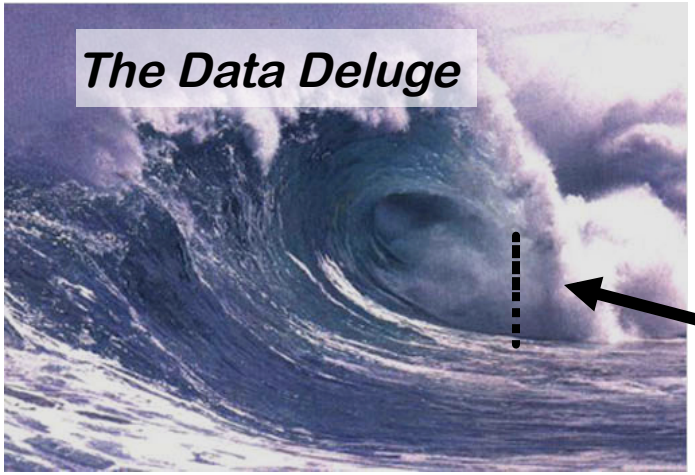
- *"All science is either physics or stamp collecting."*
  - Ernest Rutherford, British chemist & physicist (1871 - 1937)
    [J. B. Birks "Rutherford at Manchester" (1962)]
- That is, from few data, lots of thinking
- … to LOTS OF DATA and ANALYSIS

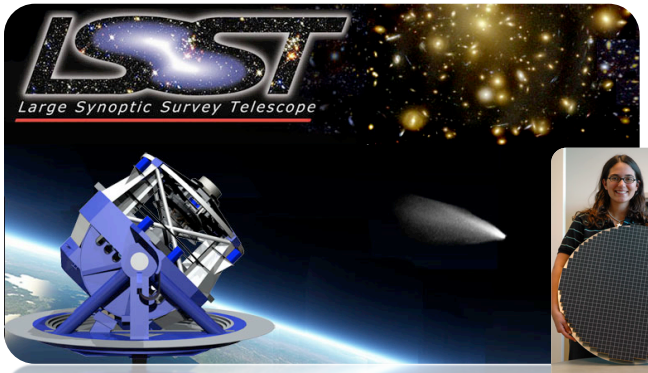
➔ "Data-driven" scientific discovery!

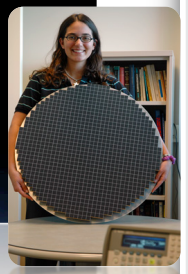
4th paradigm, in addition to hypothesis-driven science





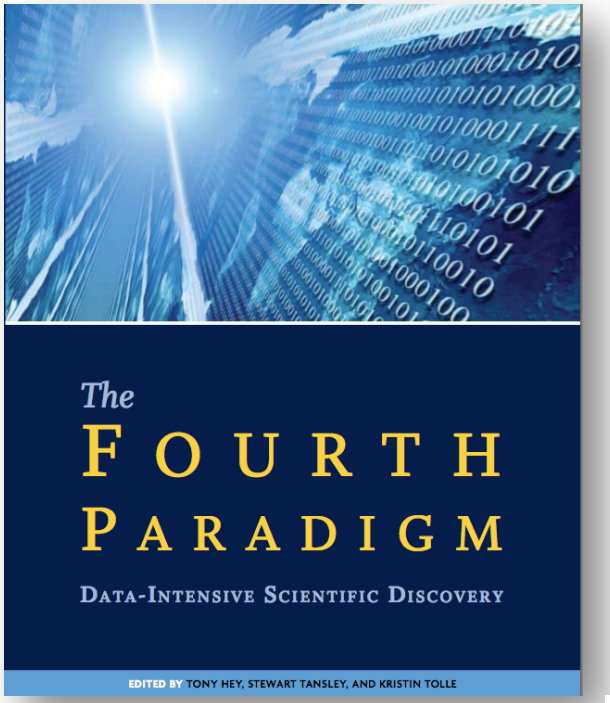






You are here …

The 4th Paradigm

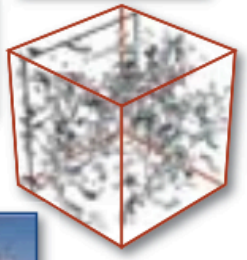


Science Paradigms

- Thousand years ago:
  science was **empirical**
  describing natural phenomena
- Last few hundred years:
  theoretical branch
  using models, generalizations

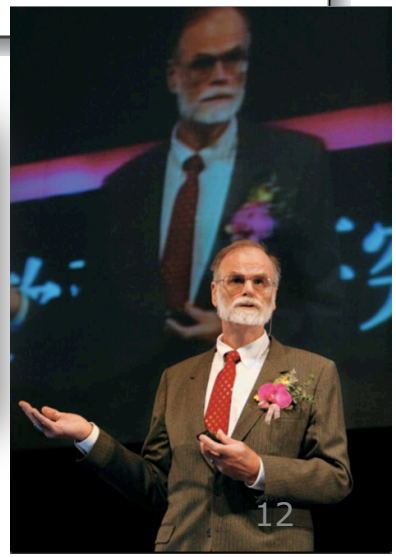
$$\left(\frac{\dot{a}}{a}\right)^2 = \frac{4\pi G\rho}{3} - K\frac{c^2}{a^2}$$

- Last few decades:
  a **computational** branch
  simulating complex phenomena
- Today: **data exploration** (eScience)
  unify theory, experiment, and simulation
  - Data captured by instruments
    or generated by simulator
  - Processed by software
  - Information/knowledge stored in computer
  - Scientist analyzes database/files
    using data management and statistics

Jim Gray on eScience:
A Transformed Scientific Method

Based on the transcript of a talk given by Jim Gray to the NRC-CSTB[1] in Mountain View, CA, on January 11, 2007[2]



Some Characteristics of Data in Databases

Data is persistent
- One or more applications use the same data
- Data stored between applications

Data often too large to easily manage in-memory
- DBMSs handle this for free
- Manually handling data (files) is usually *ad hoc* (each app. does it differently) and can be inefficient

Data may be very large (business, government, science, …)
- Library of congress > 20 terabytes of print
- Amazon.com: > 42 terabytes of data
- Youtube: > 45 terabytes of video
- AT&T: > 323 terabytes of call records
- National Energy Research Scientific Computing Center: > 2.8 petabytes

* 1 terabyte ≈ 1,000,000,000,000 bytes

* 1 petabyte ≈ 1,000,000,000,000,000 bytes (and there is talk about exabytes at DOE)

Also: Data(bases) can be Yummy!

Exploits of a Mom http://xkcd.com/327/