Data Science & Analytics

Hiren Deliwala & Dr. Jian ZHANG

Introduction to Database Systems

Topics We’ll Cover

- Databases
- Data Modeling
- SQL
- Data Analytics Using Excel
- Data Warehouse
- Business Intelligence
- Reporting & Querying
- OLAP and Multidimensional Analysis
- Statistical Analysis

What you want to learn…

Visualizations

More stuff about database

How analytics used in Medicine

How analytics used in Gaming

Data Mining

Big Data

Agenda

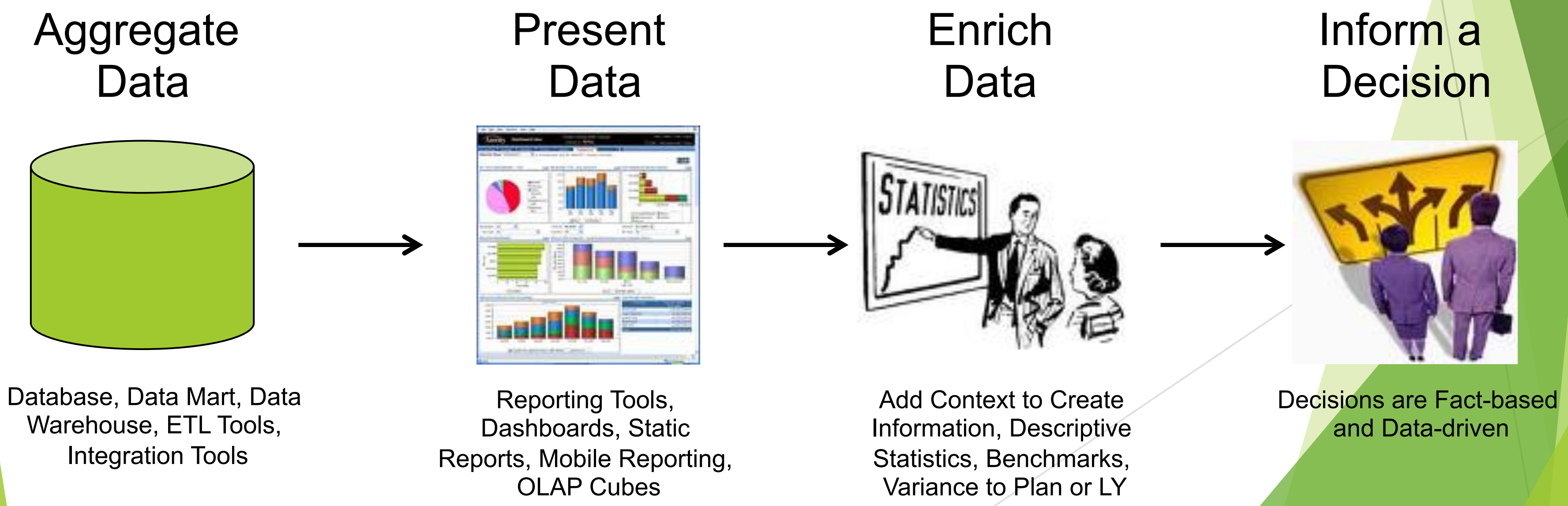
- Analytics Architecture
- Analytics Theory
- Data Analytics in Action
- Introduction to Database Systems

Business Analytics & Business Intelligence

everybody has an opinion, but nobody knows, and you shouldn’t care.

Business Intelligence

Business Intelligence enables the business to make intelligent, fact-based decisions



Aggregate Data

Database, Data Mart, Data Warehouse, ETL Tools, Integration Tools

Present Data

Reporting Tools, Dashboards, Static Reports, Mobile Reporting, OLAP Cubes

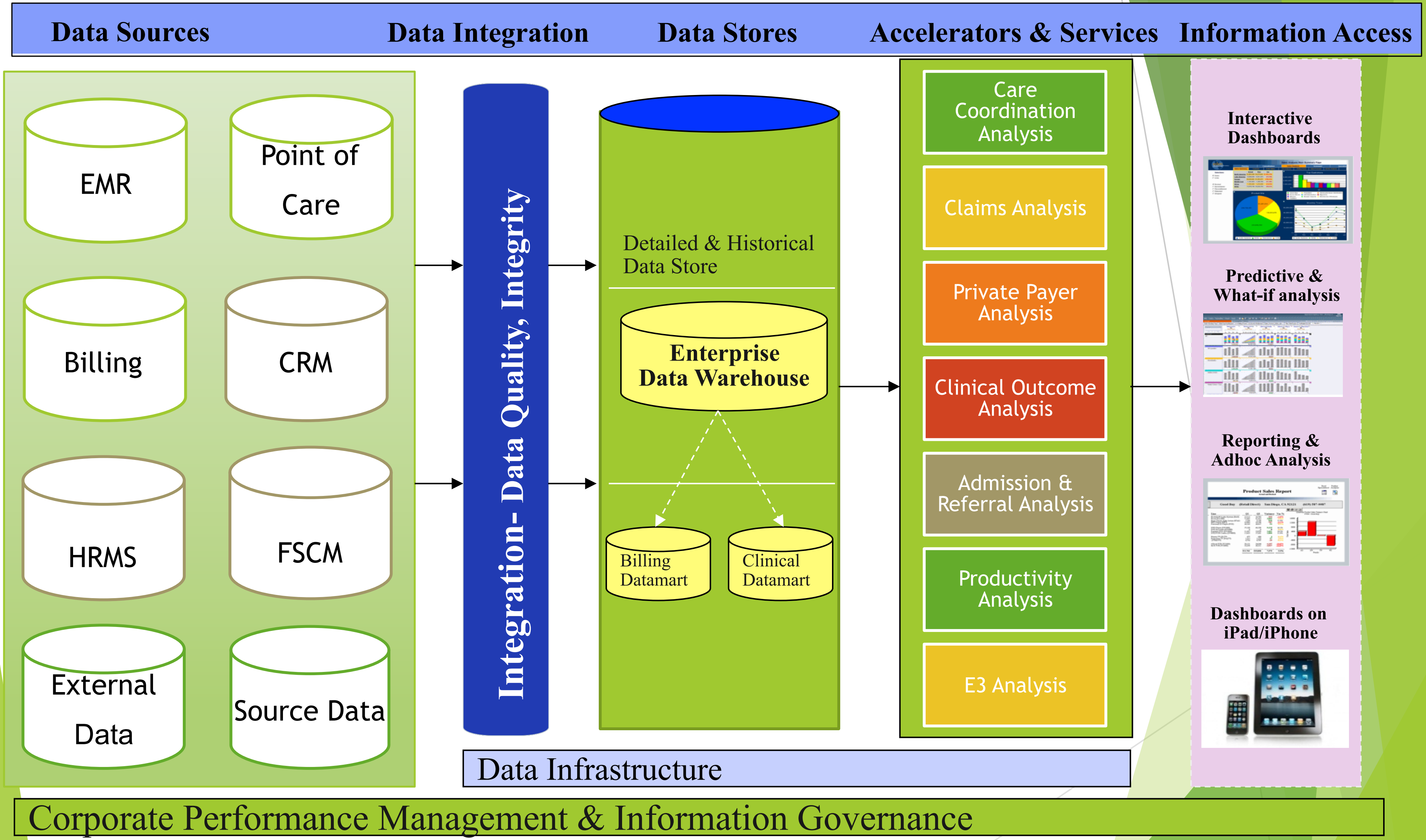
Enrich Data

Add Context to Create Information, Descriptive Statistics, Benchmarks, Variance to Plan or LY

Inform a Decision

Decisions are Fact-based and Data-driven

Analytics Solution Architecture



Data Sources

Data Integration

Data Stores

Accelerators & Services

Information Access

EMR

Point of Care

Billing

CRM

HRMS

FSCM

External Data

Source Data

Integration- Data Quality, Integrity

Detailed & Historical Data Store

Enterprise Data Warehouse

Billing Datamart

Clinical Datamart

Care Coordination Analysis

Claims Analysis

Private Payer Analysis

Clinical Outcome Analysis

Admission & Referral Analysis

Productivity Analysis

E3 Analysis

Interactive Dashboards

Predictive & What-if analysis

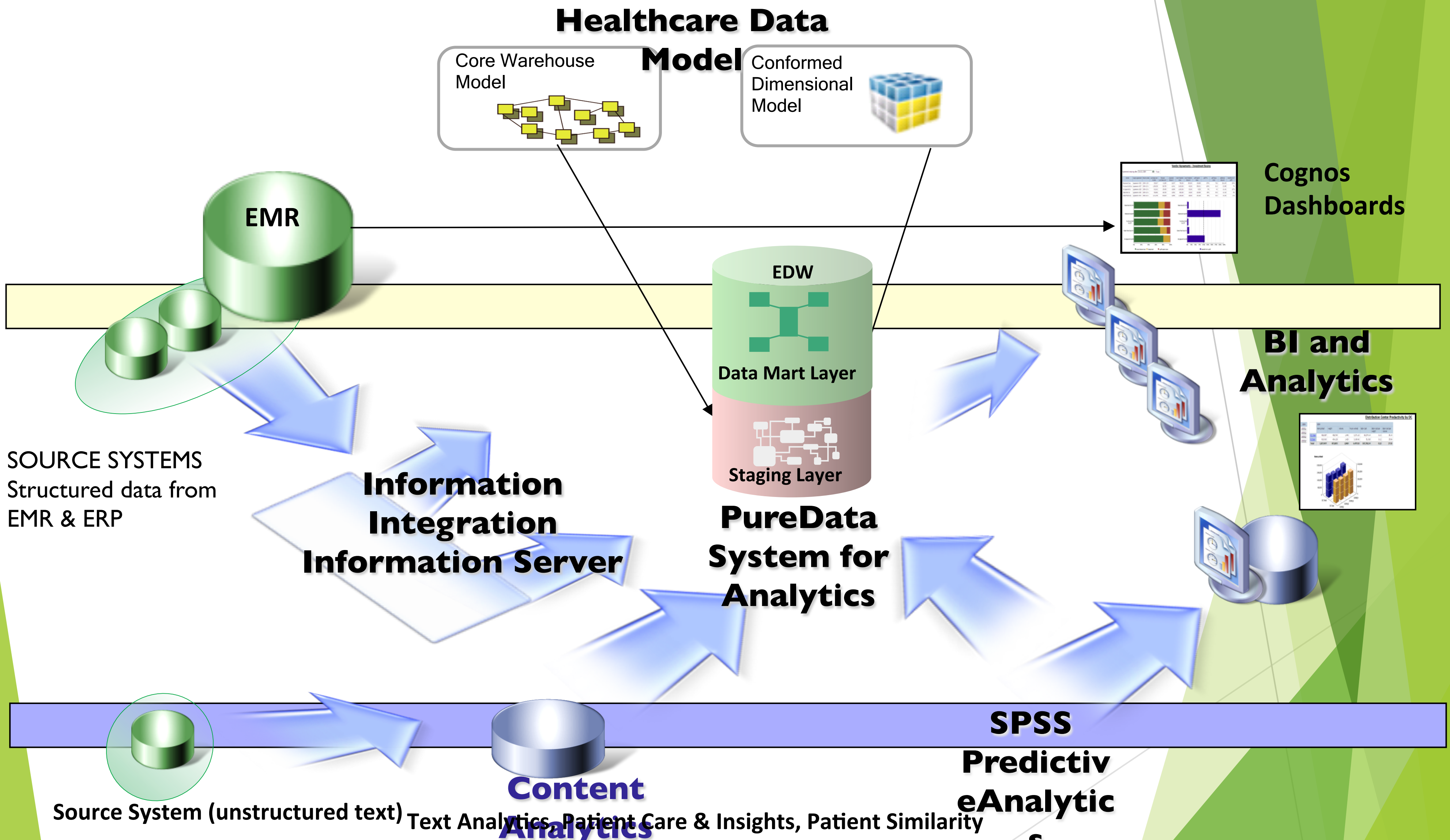
Reporting & Adhoc Analysis

Dashboards on iPad/iPhone

Data Infrastructure

Corporate Performance Management & Information Governance

Business Intelligence/Data Warehouse Architecture



Healthcare Data Model

Core Warehouse Model

Conformed Dimensional Model

EMR

Cognos Dashboards

EDW

Data Mart Layer

Staging Layer

BI and Analytics

SOURCE SYSTEMS
Structured data from
EMR & ERP

Information Integration Information Server

PureData System for Analytics

SPSS Predictiv eAnalytic s

Content Analytics

Source System (unstructured text)

Text Analytics, Patient Care & Insights, Patient Similarity

Data Governance, Regulatory Compliance, Security
(Processes in scope, products out of scope)

Analytics - Four Key Disciplines

1. Information Management
2. Statistics
3. Information Delivery
4. High Performance Computing

Information Management

- Aggregating, standardizing, restructure, and preparing the rising quantities of data for analysis
- What is it?
- Where is it?
- How good is it?
- Is there enough of it?
- Is it ready for analysis?

Information Management

1. Data Governance
2. Data Provisioning
3. Data Aggregation
4. Data Enrichment
5. Data Structuring
6. Data Quality
7. Data Integration

Database Systems

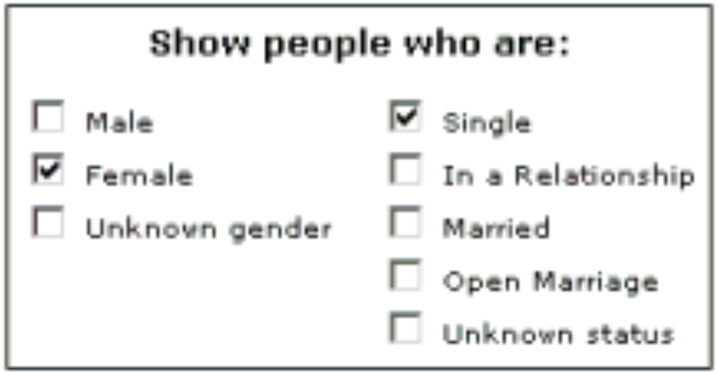
What Is a Database *System*?

- Database:
  a very large, integrated collection of data.
- Models a real-world *enterprise*
  - Entities (e.g., teams, games)
  - Relationships (e.g., Denver Broncos *are playing in* The Superbowl)
  - More recently, also includes active components , often called “business logic”. (e.g., the BCS ranking system)

- A *Database Management System (DBMS)* is a software system designed to store, manage, and facilitate access to databases.

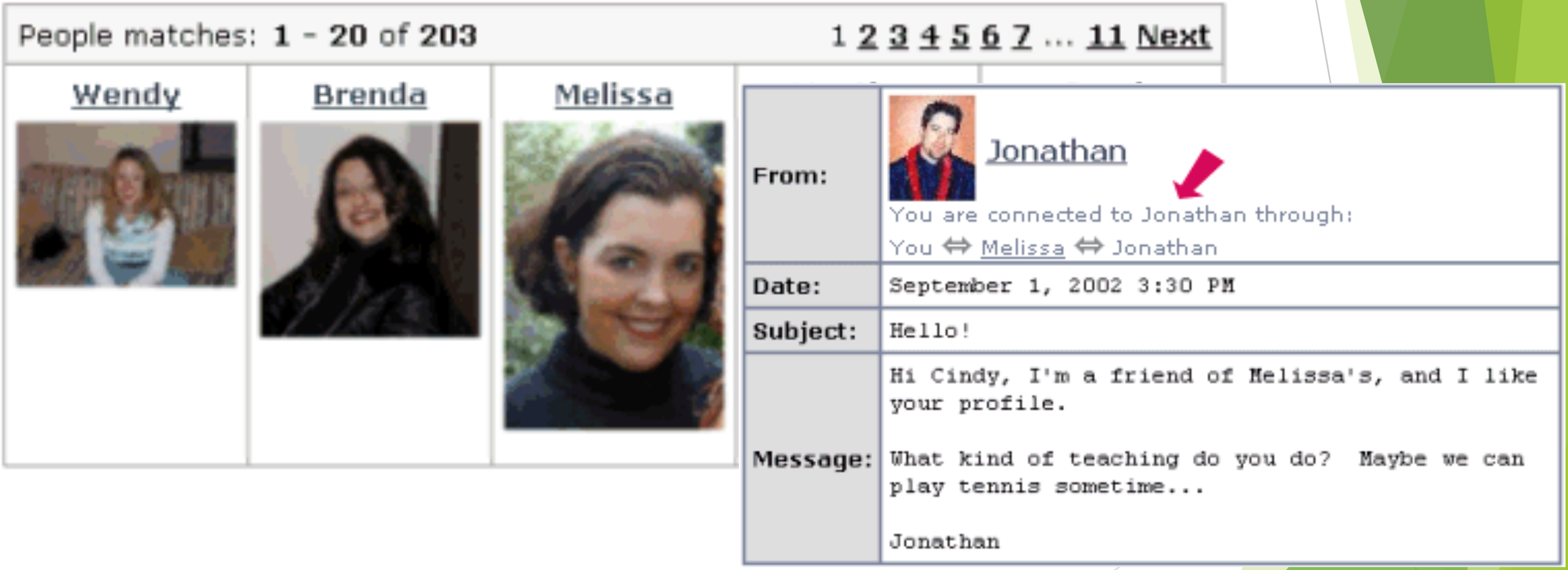
Database Systems: Then

Database Systems: Today



Show people who are:

- □ Male
- ☑ Female
- □ Unknown gender
- ☑ Single
- □ In a Relationship
- □ Married
- □ Open Marriage
- □ Unknown status

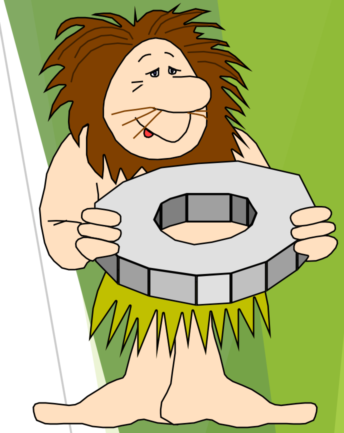


People matches: **1 - 20 of 203** 1 2 3 4 5 6 7 ... 11 Next

From:	Jonathan You are connected to Jonathan through: You ⇔ Melissa ⇔ Jonathan
Date:	September 1, 2002 3:30 PM
Subject:	Hello!
Message:	Hi Cindy, I'm a friend of Melissa's, and I like your profile. What kind of teaching do you do? Maybe we can play tennis sometime... Jonathan

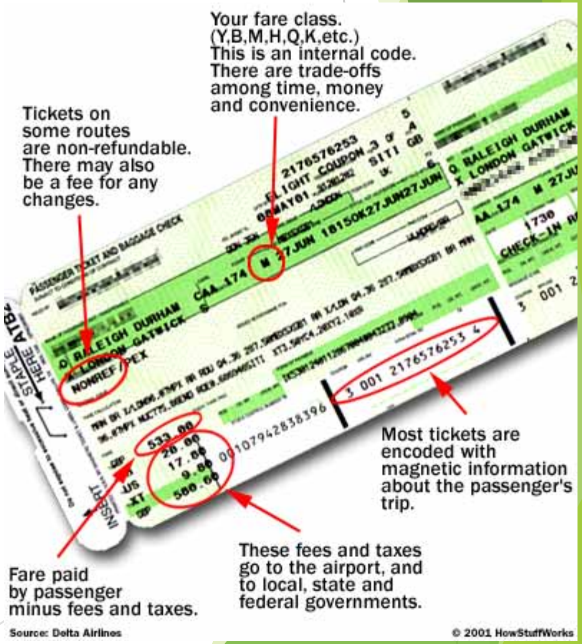
From Friendster.com o

Database Management System (DBMS)



- Collection of interrelated data
- Set of programs to access the data
- DBMS contains information about a particular enterprise
- DBMS provides an environment that is both *convenient* and *efficient* to use.
- Database Applications:
  - Banking: all transactions
  - Airlines: reservations, schedules
  - Universities: registration, grades
  - Sales: customers, products, purchases
  - Manufacturing: production, inventory, orders, supply chain
  - Human resources: employee records, salaries, tax deductions
- Databases touch all aspects of our lives

Databases you may use



Your fare class. (Y,B,M,H,Q,K,etc.) This is an internal code. There are trade-offs among time, money and convenience.

Tickets on some routes are non-refundable. There may also be a fee for any changes.

Most tickets are encoded with magnetic information about the passenger's trip.

These fees and taxes go to the airport, and to local, state and federal governments.

Fare paid by passenger minus fees and taxes.

Source: Delta Airlines

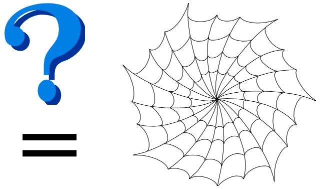
© 2001 HowStuffWorks

Purpose of Database System

- In the early days, database applications were built on top of file systems
- Drawbacks of using file systems to store data:
  - Data redundancy and inconsistency
    - Multiple file formats, duplication of information in different files
  - Difficulty in accessing data
    - Need to write a new program to carry out each new task
  - Data isolation — multiple files and formats
  - Integrity problems
    - Integrity constraints (e.g. account balance > 0) become part of program code
    - Hard to add new constraints or change existing ones

Purpose of Database Systems (Cont.)

- Drawbacks of using file systems (cont.)
  - Atomicity of updates
    - Failures may leave database in an inconsistent state with partial updates carried out
    - E.g. transfer of funds from one account to another should either complete or not happen at all
  - Concurrent access by multiple users
    - Concurrent accessed needed for performance
    - Uncontrolled concurrent accesses can lead to inconsistencies
      - E.g. two people reading a balance and updating it at the same time
  - Security problems
- Database systems offer solutions to all the above problems



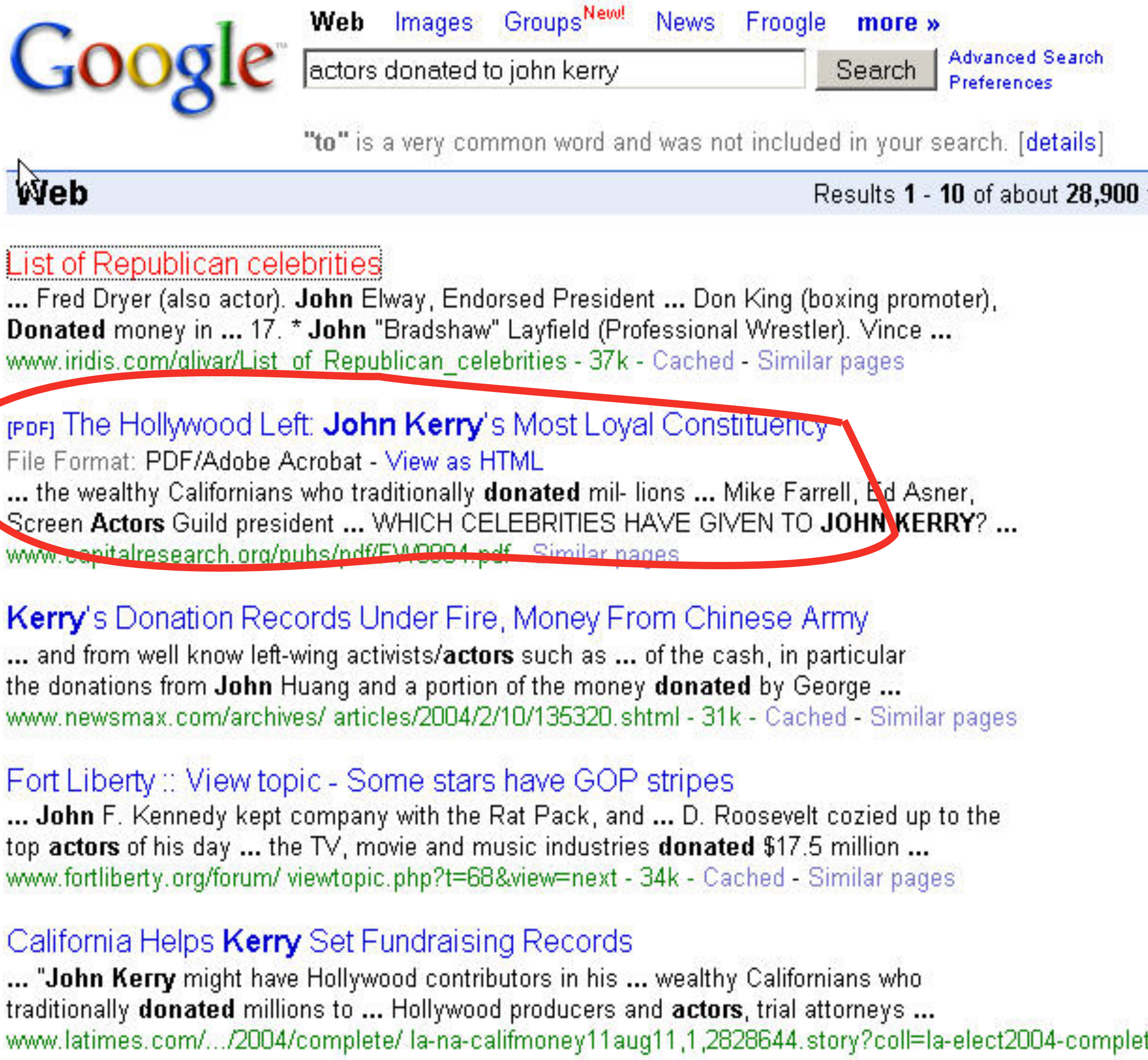
Is the WWW a DBMS?

- Fairly sophisticated search available
  - crawler *indexes* pages on the web
  - Keyword-based search for pages
- But, currently
  - data is mostly unstructured and untyped
  - search only:
    - can’t modify the data
    - can’t get summaries, complex combinations of data
  - few guarantees provided for freshness of data, consistency across data items, fault tolerance, …
  - Web sites typically have a DBMS in the background to provide these functions.
- The picture is changing
  - New standards e.g., XML, Semantic Web can help data modeling
  - Research groups (e.g., at Berkeley) are working on providing some of this functionality *across multiple web sites*.

“Search” vs. Query

- What if you wanted to find out which actors donated to John Kerry’s presidential campaign?

- Try “actors donated to john kerry” in your favorite search engine.



actors donated to john kerry

"to" is a very common word and was not included in your search. [details]

Web Results **1** - **10** of about **28,900**

List of Republican celebrities
... Fred Dryer (also actor). **John** Elway, Endorsed President ... Don King (boxing promoter), **Donated** money in ... 17. * **John** "Bradshaw" Layfield (Professional Wrestler). Vince ...
www.iridis.com/glivar/List_of_Republican_celebrities - 37k - Cached - Similar pages

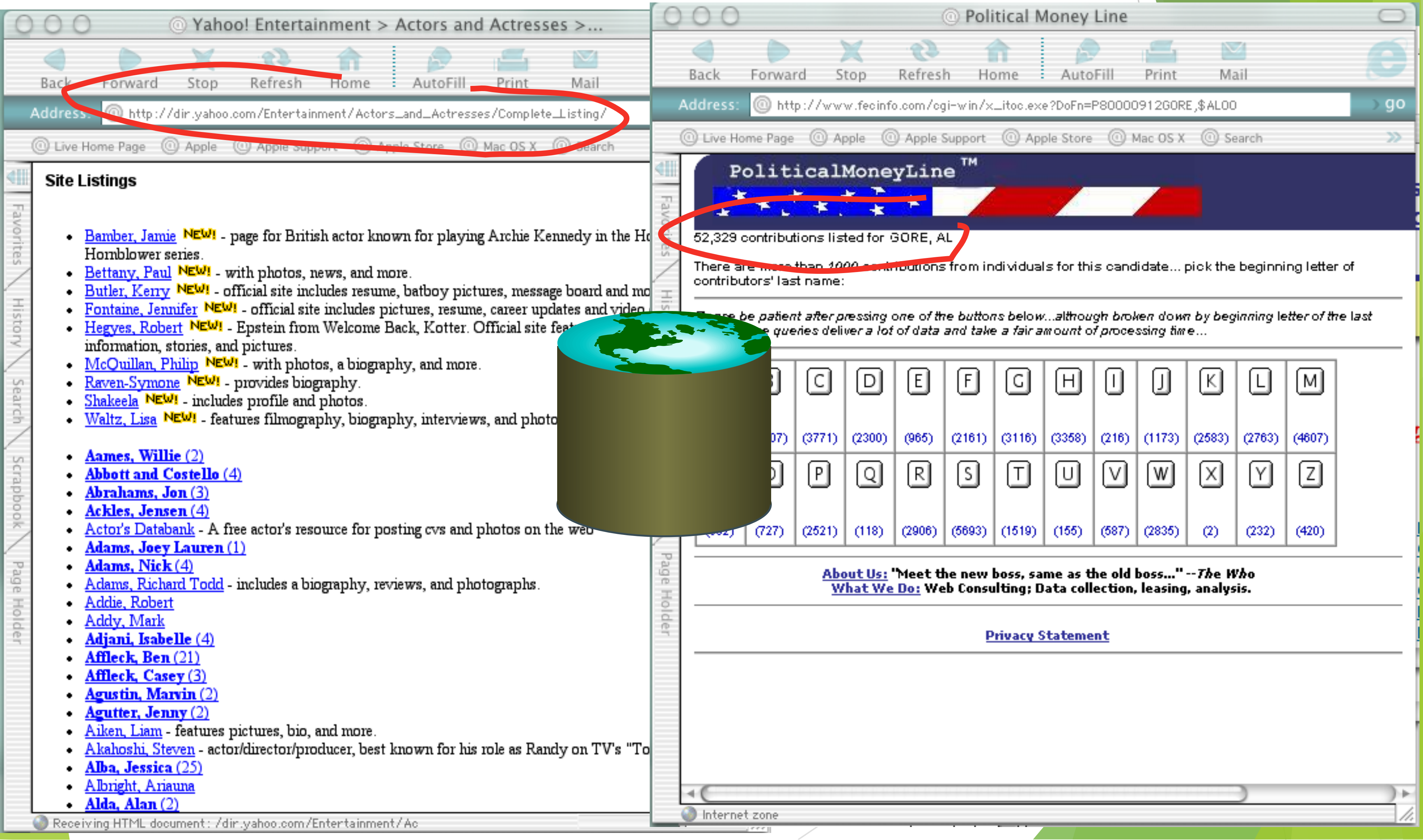
[PDF] The Hollywood Left: **John Kerry**'s Most Loyal Constituency
File Format: PDF/Adobe Acrobat - View as HTML
... the wealthy Californians who traditionally **donated** mil- lions ... Mike Farrell, Ed Asner, Screen **Actors** Guild president ... WHICH CELEBRITIES HAVE GIVEN TO **JOHN KERRY**? ...
www.capitalresearch.org/pubs/pdf/FW0904.pdf - Similar pages

Kerry's Donation Records Under Fire, Money From Chinese Army
... and from well know left-wing activists/**actors** such as ... of the cash, in particular the donations from **John** Huang and a portion of the money **donated** by George ...
www.newsmax.com/archives/ articles/2004/2/10/135320.shtml - 31k - Cached - Similar pages

Fort Liberty :: View topic - Some stars have GOP stripes
... **John** F. Kennedy kept company with the Rat Pack, and ... D. Roosevelt cozied up to the top **actors** of his day ... the TV, movie and music industries **donated** $17.5 million ...
www.fortliberty.org/forum/ viewtopic.php?t=68&view=next - 34k - Cached - Similar pages

California Helps **Kerry** Set Fundraising Records
... "**John Kerry** might have Hollywood contributors in his ... wealthy Californians who traditionally **donated** millions to ... Hollywood producers and **actors**, trial attorneys ...
www.latimes.com/.../2004/complete/ la-na-califmoney11aug11,1,2828644.story?coll=la-elect2004-complet

A “Database Query” Approach



http://dir.yahoo.com/Entertainment/Actors_and_Actresses/Complete_Listing/

Site Listings

- Bamber, Jamie NEW! - page for British actor known for playing Archie Kennedy in the Ho Hornblower series.
- Bettany, Paul NEW! - with photos, news, and more.
- Butler, Kerry NEW! - official site includes resume, batboy pictures, message board and mo
- Fontaine, Jennifer NEW! - official site includes pictures, resume, career updates and video
- Hegyes, Robert NEW! - Epstein from Welcome Back, Kotter. Official site feat information, stories, and pictures.
- McQuillan, Philip NEW! - with photos, a biography, and more.
- Raven-Symone NEW! - provides biography.
- Shakeela NEW! - includes profile and photos.
- Waltz, Lisa NEW! - features filmography, biography, interviews, and photo

- **Aames, Willie** (2)
- **Abbott and Costello** (4)
- **Abrahams, Jon** (3)
- **Ackles, Jensen** (4)
- Actor's Databank - A free actor's resource for posting cvs and photos on the web
- **Adams, Joey Lauren** (1)
- **Adams, Nick** (4)
- Adams, Richard Todd - includes a biography, reviews, and photographs.
- Addie, Robert
- Addy, Mark
- **Adjani, Isabelle** (4)
- **Affleck, Ben** (21)
- **Affleck, Casey** (3)
- **Agustin, Marvin** (2)
- **Agutter, Jenny** (2)
- Aiken, Liam - features pictures, bio, and more.
- Akahoshi, Steven - actor/director/producer, best known for his role as Randy on TV's "To
- **Alba, Jessica** (25)
- Albright, Ariauna
- **Alda, Alan** (2)

Receiving HTML document: /dir.yahoo.com/Entertainment/Ac

http://www.fecinfo.com/cgi-win/x_itoc.exe?DoFn=P80000912GORE,$AL00

52,329 contributions listed for GORE, AL

There are more than 1000 contributions from individuals for this candidate... pick the beginning letter of contributors' last name:

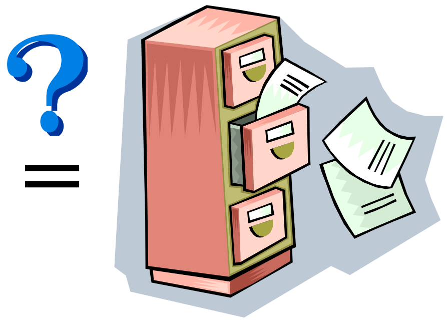
[illegible] be patient after pressing one of the buttons below...although broken down by beginning letter of the last [illegible] queries deliver a lot of data and take a fair amount of processing time...

[illegible]	B	C	D	E	F	G	H	I	J	K	L	M
[illegible]	[illegible]07)	(3771)	(2300)	(965)	(2161)	(3116)	(3358)	(216)	(1173)	(2583)	(2763)	(4607)
[illegible]	O	P	Q	R	S	T	U	V	W	X	Y	Z
[illegible]	(727)	(2521)	(118)	(2906)	(5693)	(1519)	(155)	(587)	(2835)	(2)	(232)	(420)

About Us: "Meet the new boss, same as the old boss..." --*The Who*
What We Do: Web Consulting; Data collection, leasing, analysis.

Privacy Statement

Internet zone



Is a File System a DBMS?

- Thought Experiment 1:
  - You and your project partner are editing the same file.
  - You both save it at the same time.
  - Whose changes survive?

A) Yours B) Partner's C) Both D) Neither E) ???

- **Thought Experiment 2:**
  - You're updating a file.
  - The power goes out.
  - Which of your changes survive?

A) All B) None C) All Since Last Save D) ???

Q: How do you write programs over a subsystem when it promises you only "???" ?

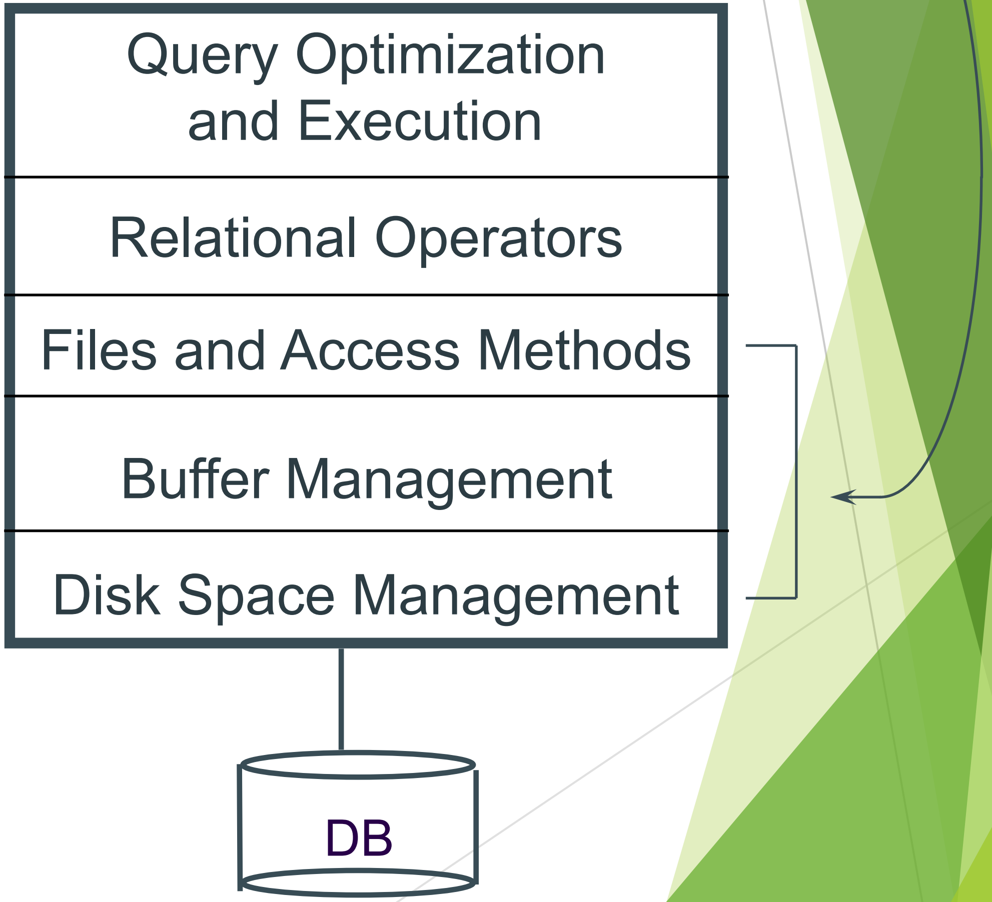
A: Very, very carefully!!

Current Commercial Outlook

- A major part of the software industry:
  - Oracle, IBM, Microsoft, Sybase
  - Netezza, Teradata, GreenPlum
  - smaller players: java-based dbms, devices, OO, …
- Well-known benchmarks (esp. TPC)
- Lots of related industries
  - data warehouse, document management, storage, backup, reporting, business intelligence, app integration
- Relational products dominant and evolving
  - adapting for extensibility (user-defined types), adding native XML support.
- Open Source coming on strong
  - MySQL, PostgreSQL, BerkeleyDB

Structure of a DBMS

- A typical DBMS has a layered architecture.
- The figure does not show the concurrency control and recovery components.
- Each database system has its own variations.



These layers must consider concurrency control and recovery

Why Study Databases??

- Shift from *computation* to *information*
  - always true for corporate computing
  - Web made this point for personal computing
  - more and more true for scientific computing
- Need for DBMS has exploded in the last years
  - Corporate: retail swipe/clickstreams, “customer relationship mgmt”, “supply chain mgmt”, “data warehouses”, etc.
  - Scientific: digital libraries, Human Genome project, NASA Mission to Planet Earth, physical sensors, grid physics network
- DBMS encompasses much of CS in a practical discipline
  - OS, languages, theory, AI, multimedia, logic
  - Yet traditional focus on real-world apps

Collecting and storing big data alone isn't enough to produce real business value. Analytics is necessary to:

1. Formulate eye-catching charts and graphs
2. Extract valuable insights from the data
3. Integrate data from internal and external sources

Companies that have large amounts of information stored in different systems should begin a big data analytics project by considering:

1. The creation of a plan for choosing and implementing big data infrastructure technologies
2. The interrelatedness of data and the amount of development work that will be needed to link various data sources
3. The ability of business intelligence and analytics vendors to help them answer business questions in big data environments

Recommended best practices for managing data analytics programs include:

1. Adopting data analysis tools based on a laundry list of their capabilities
2. Letting go entirely of "old ideas" related to data management
3. Focusing on business goals and how to use big data analytics to meet them

Instances and Schemas

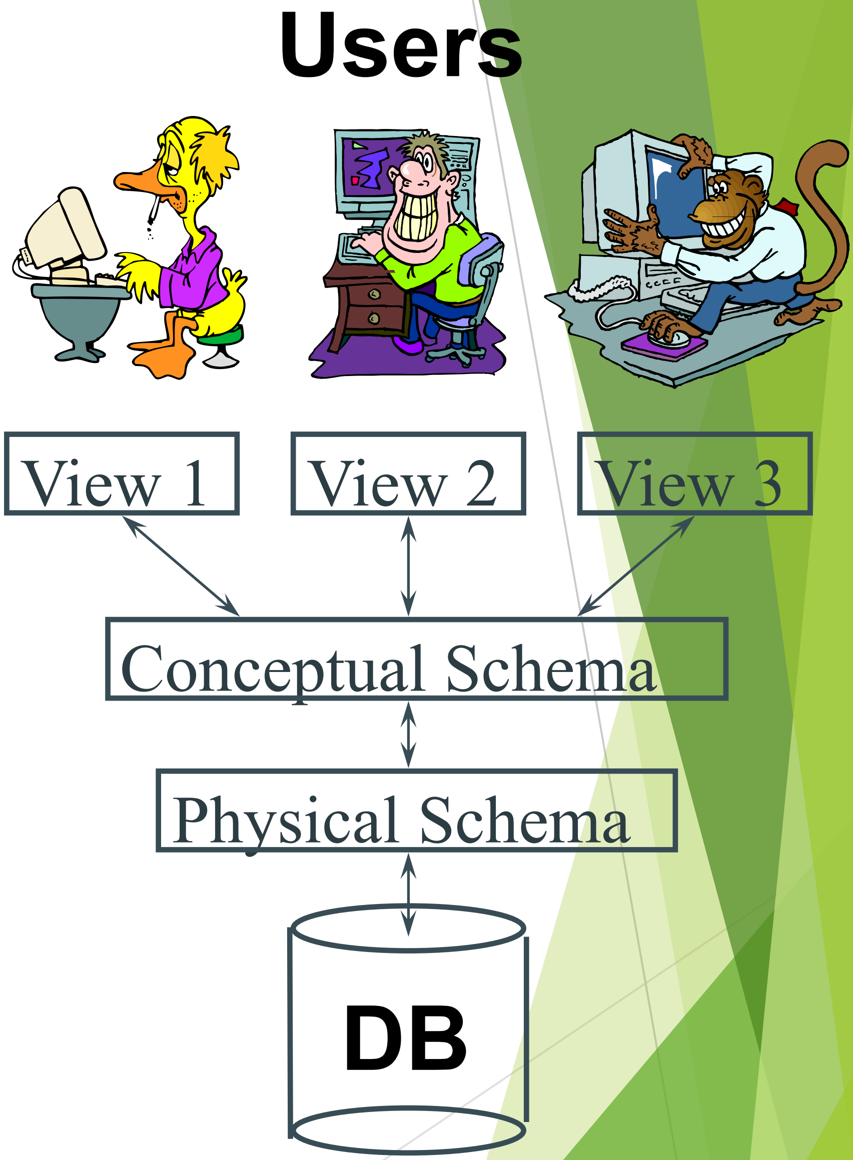
- Similar to types and variables in programming languages
- **Schema** – the logical structure of the database
  - Example: The database consists of information about a set of customers and accounts and the relationship between them)
  - Analogous to type information of a variable in a program
  - **Physical schema:** database design at the physical level
  - **Logical schema:** database design at the logical level
- **Instance** – the actual content of the database at a particular point in time
  - Analogous to the value of a variable
- **Physical Data Independence** – the ability to modify the physical schema without changing the logical schema
  - Applications depend on the logical schema
  - In general, the interfaces between the various levels and components should be well defined so that changes in some parts do not seriously influence others.

Data Models

- A collection of modeling tools for describing
  - data
  - data relationships
  - data semantics
  - data constraints
- Entity-Relationship model
- Relational model
- Other models:
  - object-oriented model
  - semi-structured data models (XML)
  - Older models: network model and hierarchical model

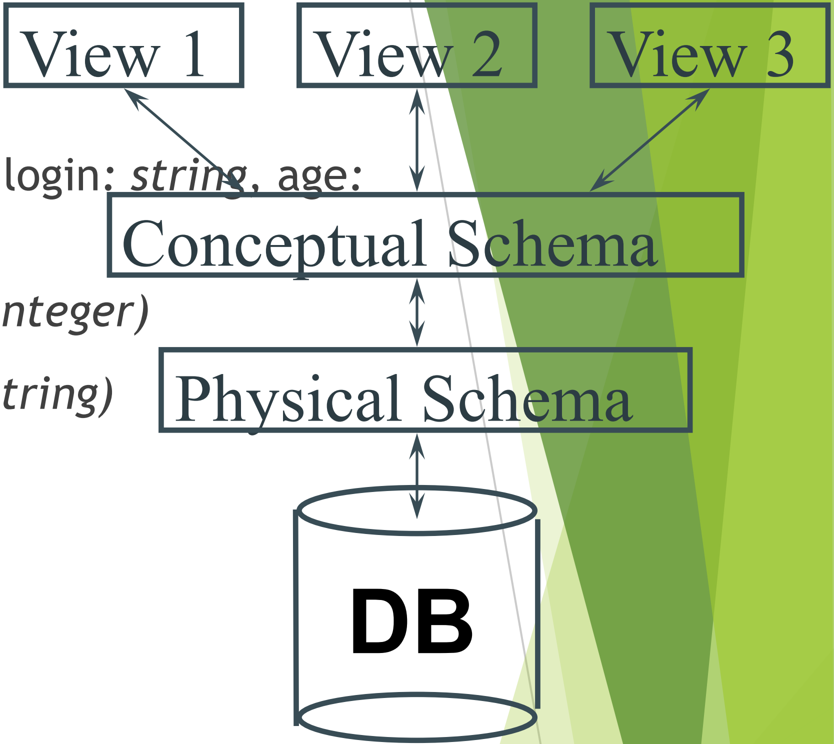
Levels of Abstraction

- Views describe how users see the data.
- Conceptual schema defines logical structure
- Physical schema describes the files and indexes used.
- (sometimes called the ANSI/SPARC model)



Example: University Database

- Conceptual schema:
  - *Students*(sid: *string*, name: *string*, login: *string*, age: *integer*, gpa:*real)*
  - *Courses*(cid: *string*, cname:*string*, credits:*integer)*
  - *Enrolled*(sid:*string*, cid:*string*, grade:*string)*
- External Schema (View):
  - *Course_info*(cid:*string*,enrollment:*integer)*
- Physical schema:
  - Relations stored as unordered files.
  - Index on first column of Students.


View 1

View 2

View 3

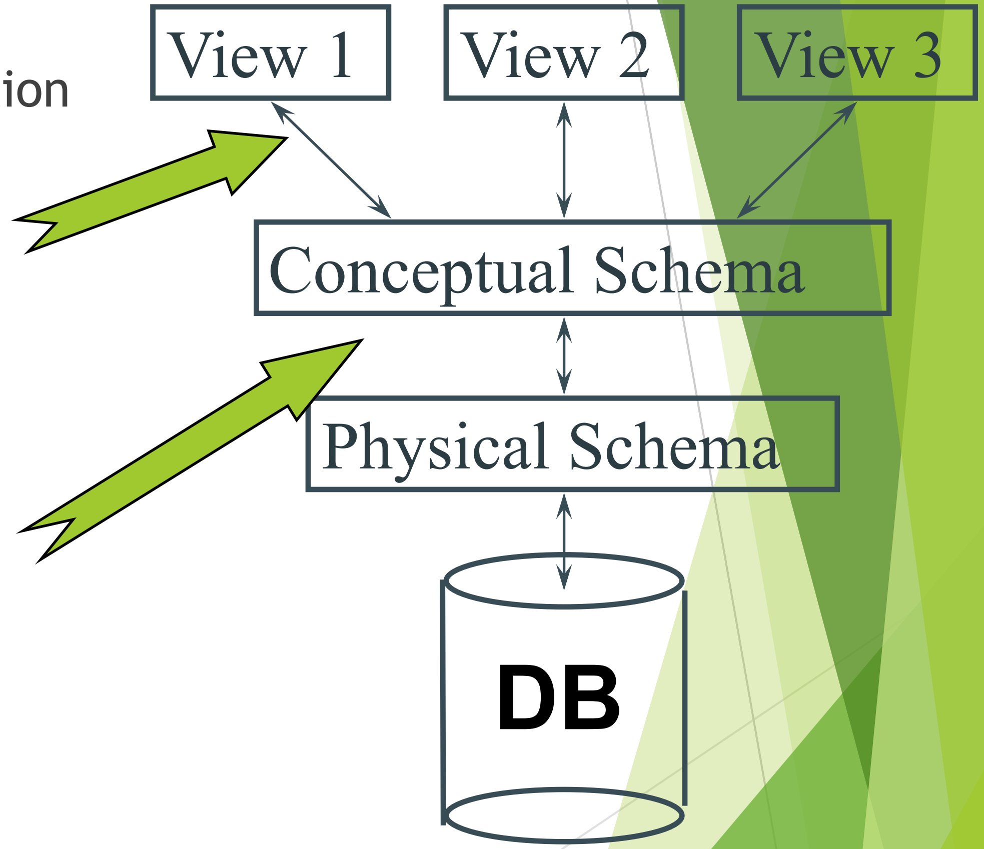
Conceptual Schema

Physical Schema

DB

Data Independence

- Applications insulated from how data is structured and stored.
- Logical data independence: Protection from changes in *logical* structure of data.
- Physical data independence: Protection from changes in *physical* structure of data.
- Q: Why are these particularly important for DBMS?



Smarter Analytics