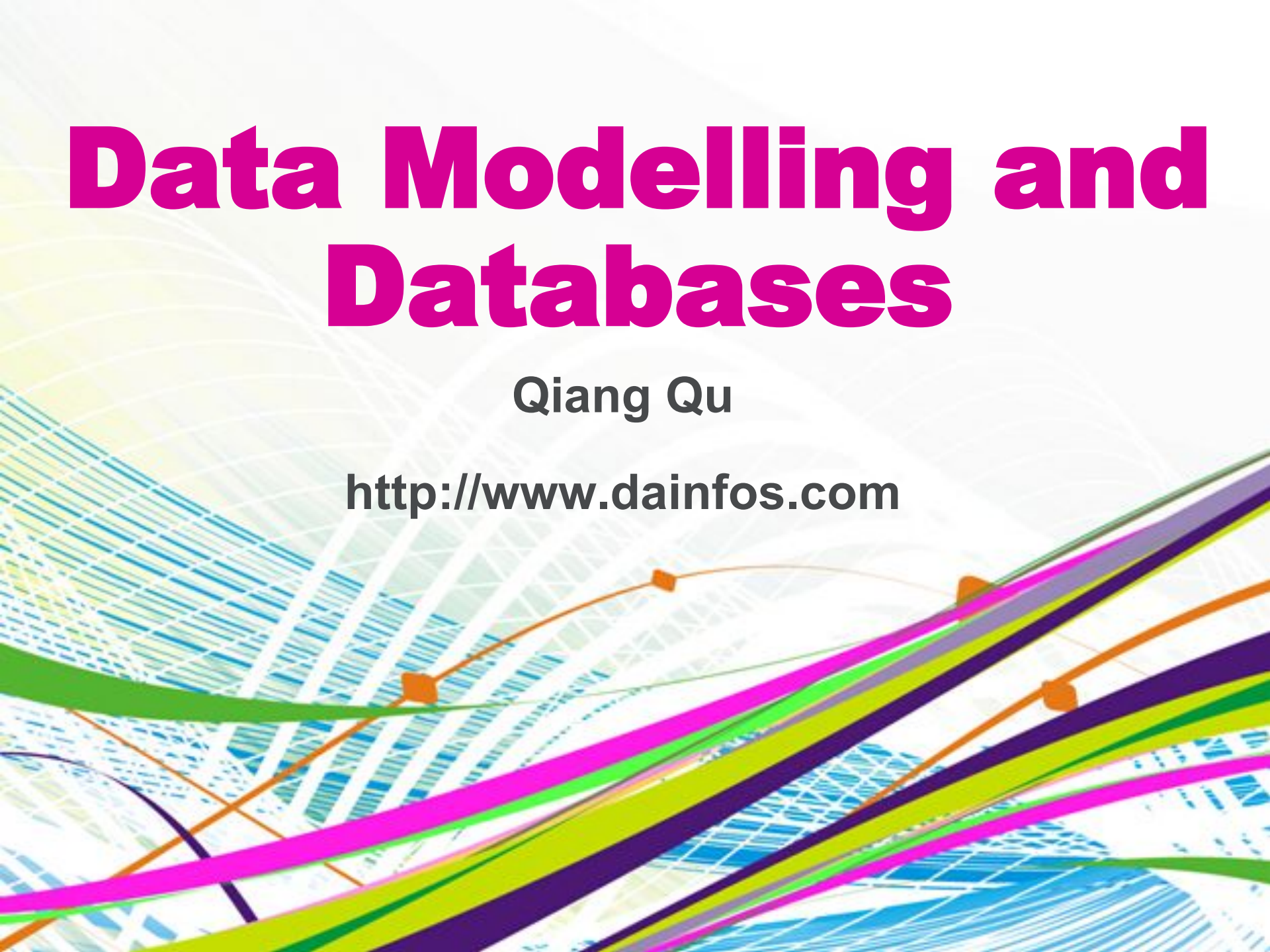
Data Modelling and Databases

Qiang Qu

http://www.dainfos.com



Course Schedule

- Lecture (Start 17 Aug 2015) - English

  Monday 9:00-10:30

- Seminar and Exercise Groups - English, Russian

  10 groups on Monday

  Rule: NO SWAP

- Logic of groups

  Grade level

  Various seminar lectures, exercises, lab sessons

  Possible language support

Stuff and Teaching Philosophy

- Primary Instructor (PI)
  Qiang Qu, qu@innopolis.ru
- Secondary Instructor (SI)
  Sadegh Nobari, nobari@innopolis.ru
- Assistant Instructor (AI)
  Jooyoung Lee, j.lee@innopolis.ru
  Waqas Nawaz, w.nawaz@innopolis.ru

- Teaching Assistant (TA)
  Rasul Tumyrkin, rasul.tumyrkin@mail.ru
  Emil Melnikov, e.melnikov@innopolis.ru
  Ais Khairullina, ais.khairullina@gmail.com

- Define curriculum + Learning by doing + Your own speed

Textbook

- A First Course in Database Systems (3rd E) by Jeffrey D. Ullman, Jennifer Widom
  http://www.amazon.com/First-Course-Database-Systems-Edition/dp/013600637X
- Slides are adapted from ETH Z.
- Reference Book:
  Concepts of Database Management (7th E) by Philip J. Pratt, Joseph J. Adamski
  http://www.amazon.com/Concepts-Database-Management-Philip-Pratt/dp/1111825912

Overview

- **How to use a database system?**
  - Data modelling (ER, theory)
  - Database programming (SQL)
- How to build a database system?
  - Query optimization
  - Transaction management
  - Internals
- What next?
  - Big Data: Data Warehousing, Data Mining
  - XML & WWW

Exercises and Exams

- Exercises and assignments
  - Handout on Monday
  - Handin before the next Monday
- Project
  - TBA in seminar groups
  - one/two, one is graded
- Written Exams – both closed book
  - Mid-term
  - Final-term

Scoring Scheme

- The 5 components matter

  Attendence: 5% (fail if in total +5 times of absence in lectures and seminars)

  Assignments: 15%

  Mid-term Exam: 20%

  Final-term Exam: 50%

  Project: 10%

INTRODUCTION

A Short History of Computing

- since 1940s: Computers for Number Crunching

  von Neuman Machine, Moore's law

- since 1990s: Magnetic storage cheaper than paper

  various technologies (tape, disk, flash, …)

- since 2000s: "The Cloud"

  sensors: most data is digitally born

  e.g., mobile phones, cars, microwaves, fitbit, …

A Short History of Computing

- since
  Moor
  (hit
- since 1990s: Magnetic storage cheaper than paper
  various technologies (disk, DRAM, SSD, PCM, …)
  continued trend of high density at same cost
- since
  va

Calculator
(+, -, *, /)

Information Hub
(store, process, communicate data)

Computer Science in Change

- Traditional Computing - automate processes

  execute a sequence of +, *, …

- Today: "Big Data"- automate experiences

  do not do the same mistake twice

  answer tough questions based on past evidence

Simple Truths

- "Power of data"
  - the more data the merrier (GB -> TB)
  - data comes from everywhere in all shapes
  - value of data often discovered later
  - data has no owner within an organization (no silos!)
- Services turn data into $
  - the more services the merrier
  - need to adapt quickly
- E.g.: Google, Amadeus, Disney, Walmart, BMW, ...
- Platforms: IBM, Oracle, MS, SAP, Google, 28msec, ...

Two Examples

- Google Translate

  translate text based on snippets of multi-lang. corpora, e.g., EU patents, translated books, Web sites, etc.

- Patients

  find patients with the same disease, summarizing features of the disease

Challenges

- Automate Experience – *NOT* Thinking!

  only works if you ask the right questions and interpret the answers correctly

  shortage of Big Data talent on job market

- Misuse of data & privacy

  owner must control usage of data

- Democratization

  Big Data opportunities in the hands of everybody

Big Data Question: Yes or No?

- Find a spouse?
- Cure for cancer?
- How to treat a cough?
- Should I give somebody a loan?
- Premium for fire insurance?
- When should my son come home?
- Which book should I read next?
- 1+1?

Vision

- Answer all questions

  Store all data and make it available and useful to all authorized people, anytime and anywhere.

- Google‘s mission statement:

Organize all the information of the world.

- Status: Technology is there (card boxes). The model is missing (labels).

Data Science: Science of Questions

- How to formulate questions?

  relational algebra

- How to organize data to answer questions?

  ER / UML, relational data model

- How to acquire data to answer questions?

  project, transactions, (much more not covered)

- How to make it efficient

  normal forms, optimization

- How to quantify error, avoid stupid questions?

  not covered in this class ☹

What is a Database System (DBMS)?

- A DBMS is a tool that helps develop and run data-intensive applications (create and maintain):

  large databases

  large data streams

- Database:

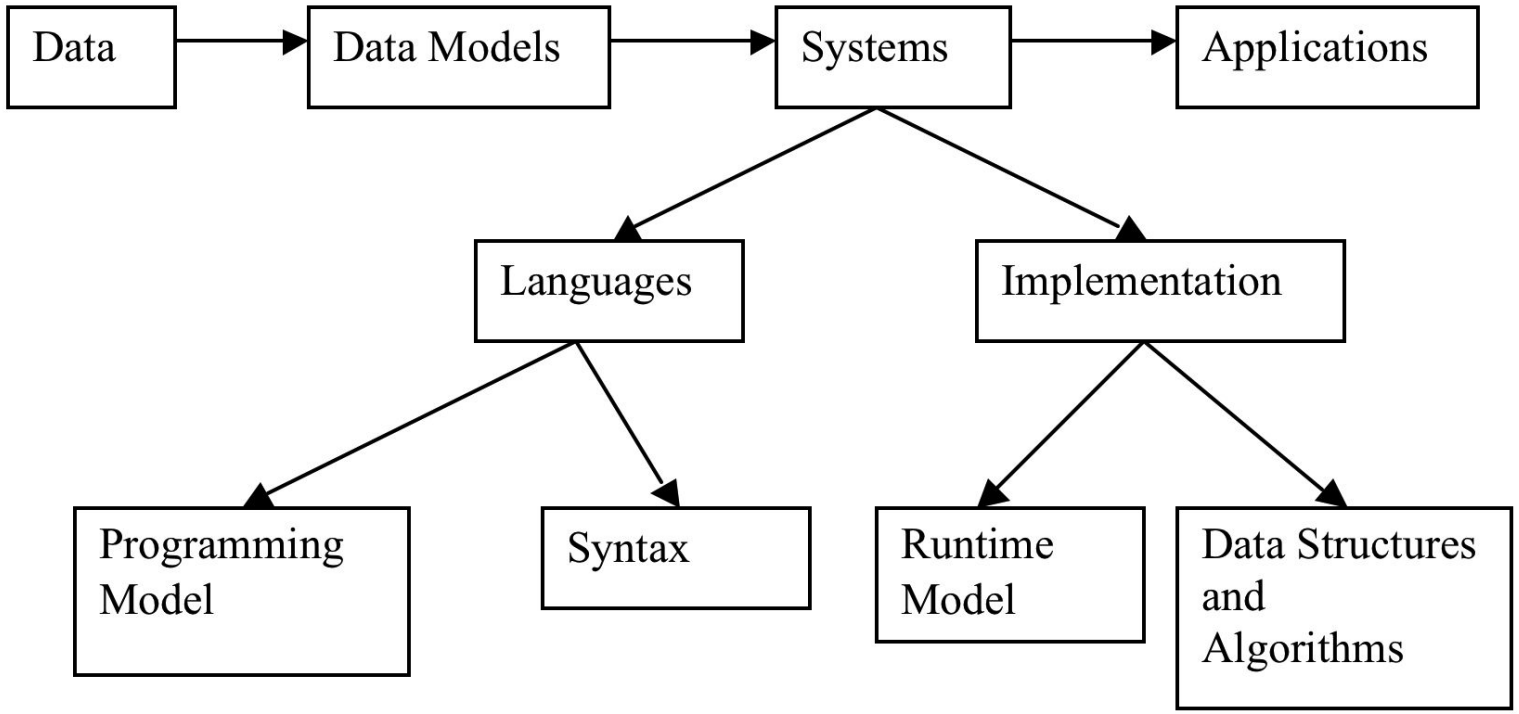
  collection of interrelated data, which

  1. represents some aspects of the real world

  2. is logically coherent with some inherent meaning

  3. has an intended group of users and applications

The Data Management Universe



DB vs. BD

- Databases
  - You know questions upfront
  - Closed world: data correct & complete
- Big Data
  - The exact opposite in all regards
- But, similar algorithms, languages, technology
  - Collect data to answer question when it is asked
  - Bridge time between event (data) and question

Data and Data Models

- Formats

  XML, serialized Java objects, binary, ...

- Structures / Models

  Tuples, hierarchies, relationships, lists, unstructured, ...

- Examples

  Lecture notes

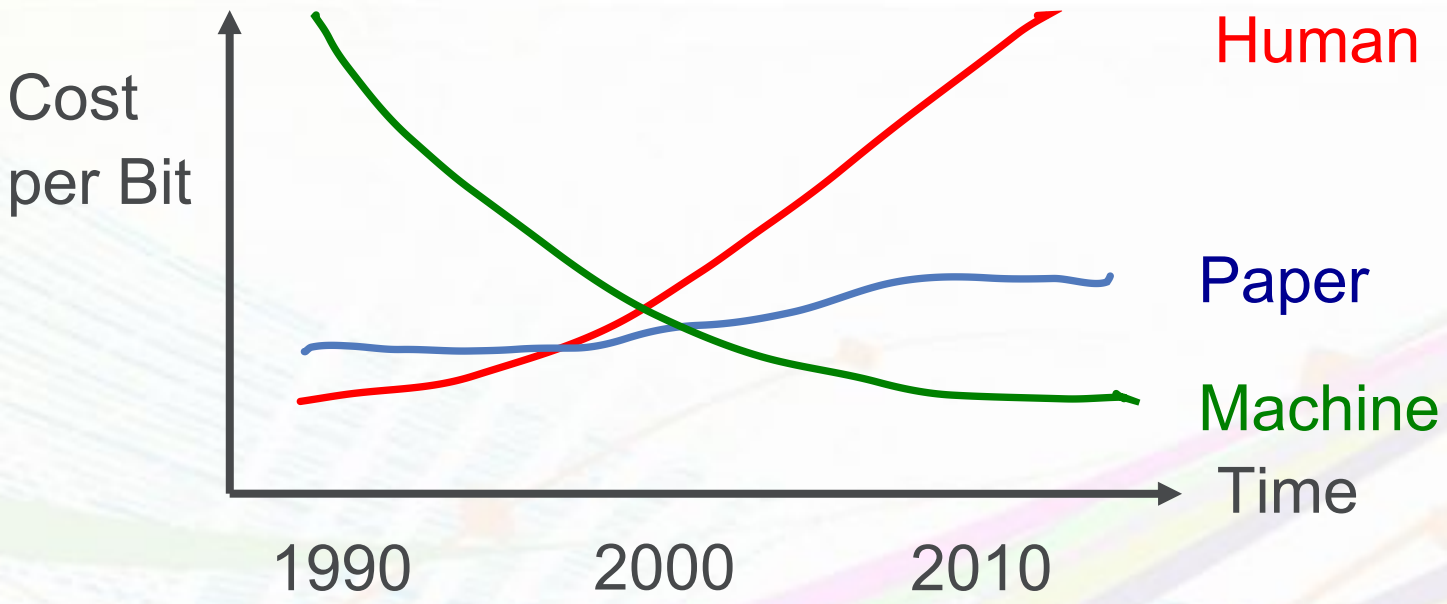
  Financial accounts

  Emotions (?): love, taste, ...

Systems

- Software platforms that store & organize data
  - File system: Windows, ...
  - Relational database systems: Postgres, Oracle, ...
  - Other database systems: OB,Sausalito, OODB, ...
  - Key/value stores: HBase, AWS S3, MongoDB, ...
  - Interpreters: JVM, .NET, ...
  - Human intelligence
- Hardware that stores & organizes data
  - HDD, SSD, main memory, ...
  - Paper
  - Human brain

Where is data stored today?



Mechanical Turk: Prices for humans going down again. How come?

Typical Applications (data / operations)

- Bank (Accounts / "Money Transfer")
- Library (Books / "Lend Book")
- Content Management System (docs, "show")
- E-Business (Catalogue, "search")
- ERP (Order, "delivery")
- Decision Support (Order, "emp of the month“)
- Facebook, Twitter, … (Friends, "post tweet")

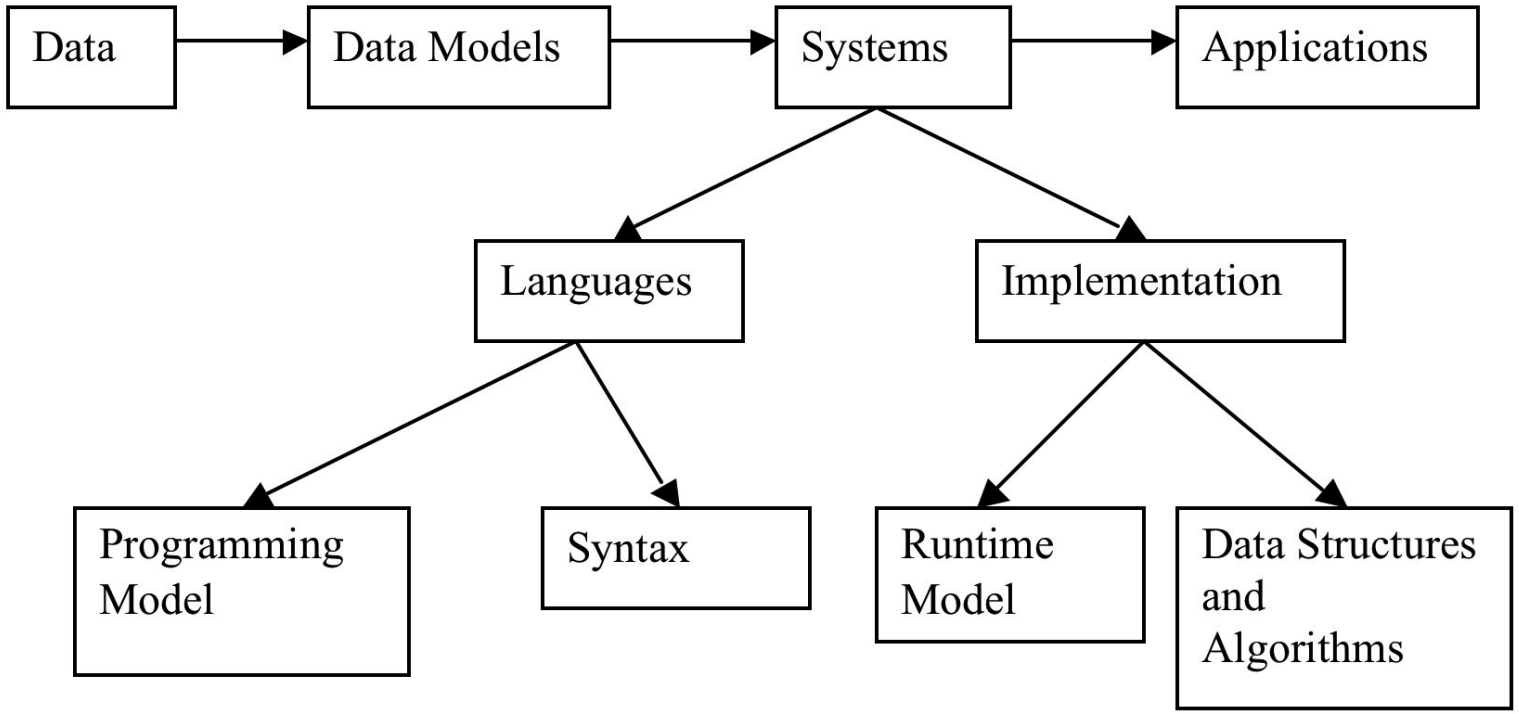
Why use a DBMS?

- Avoid redundancy and inconsistency
- Rich (declarative) access to the data
- Synchronize concurrent data access
- Recovery after system failures
- Security and privacy

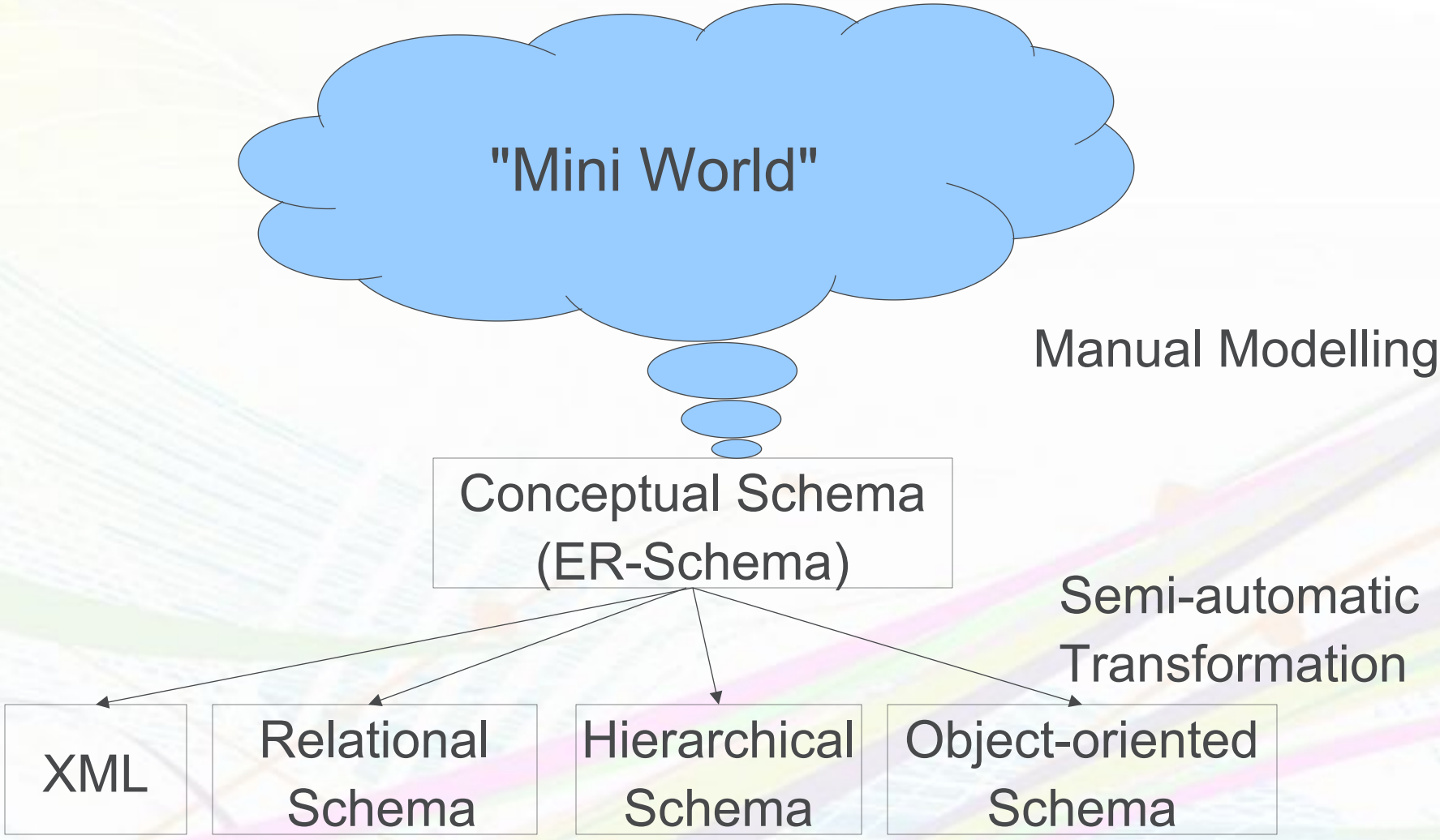
- Reduce cost and pain to do something useful

  There is always an alternative!!!

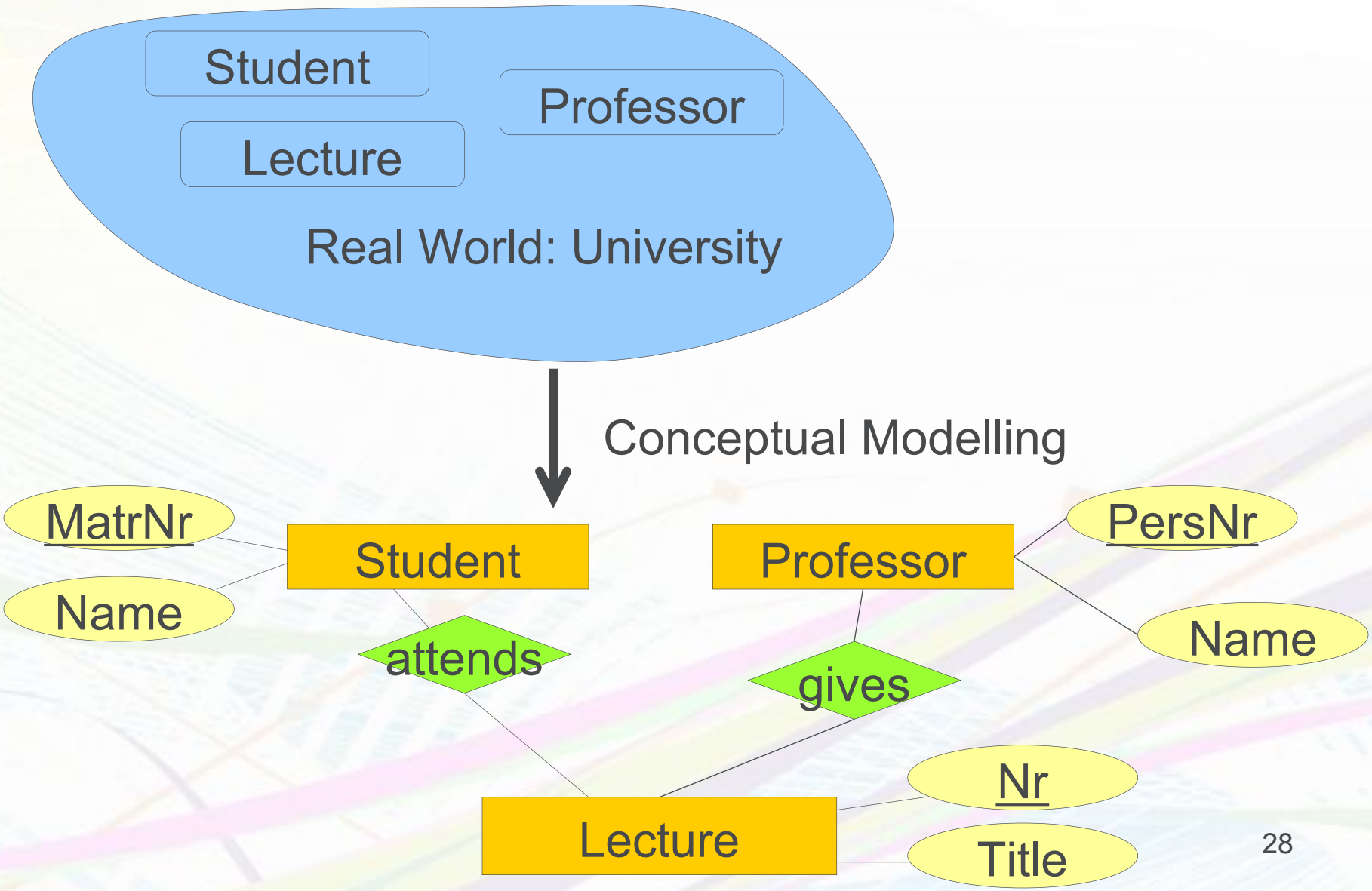
The Data Management Universe



Data Modelling



Example


Student
Professor
Lecture
Real World: University
Conceptual Modelling
MatrNr
Name
Student
attends
Professor
gives
PersNr
Name
Nr
Title
Lecture

Overview of Data Models

- Network model (e.g., CODASYL)
- Hierarchical model (IBM IMS/FastPath)
- Relational model (SQL)
- Object-oriented model (ODMG 2.0)
- Semi-structured model (XML Infoset)
- Deductive model (Datalog, Prolog)

Relational Data Model

Student	
Legi	Name
26120	Fichte
25403	Jonas
...	...

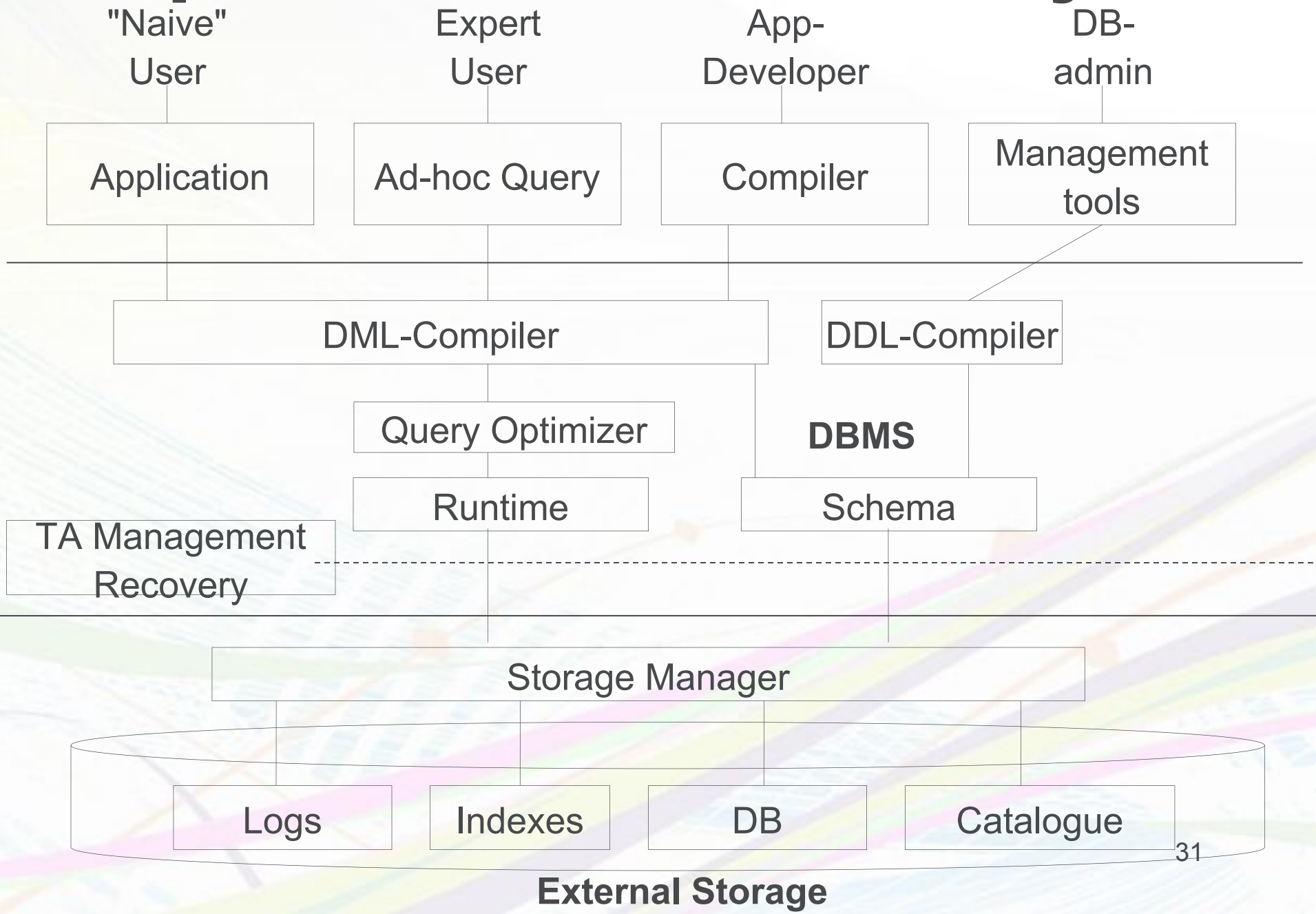
attends	
Legi	Lecture
25403	5022
26120	5001
...	...

Lecture	
Nr	Title
5001	DMD
5022	PDS
...	...

```
Select Name
From Student, attend, Lecture
Where Student.Legi= attend.Legi and
      attend.Lecture= Lecture.Nr and
      Lecture.Title = `DMS´;
```

```
Update      Lecture
   set      Title = `Data modelling and Databases´
 where      Nr = 5001;
```

Components of a Database System

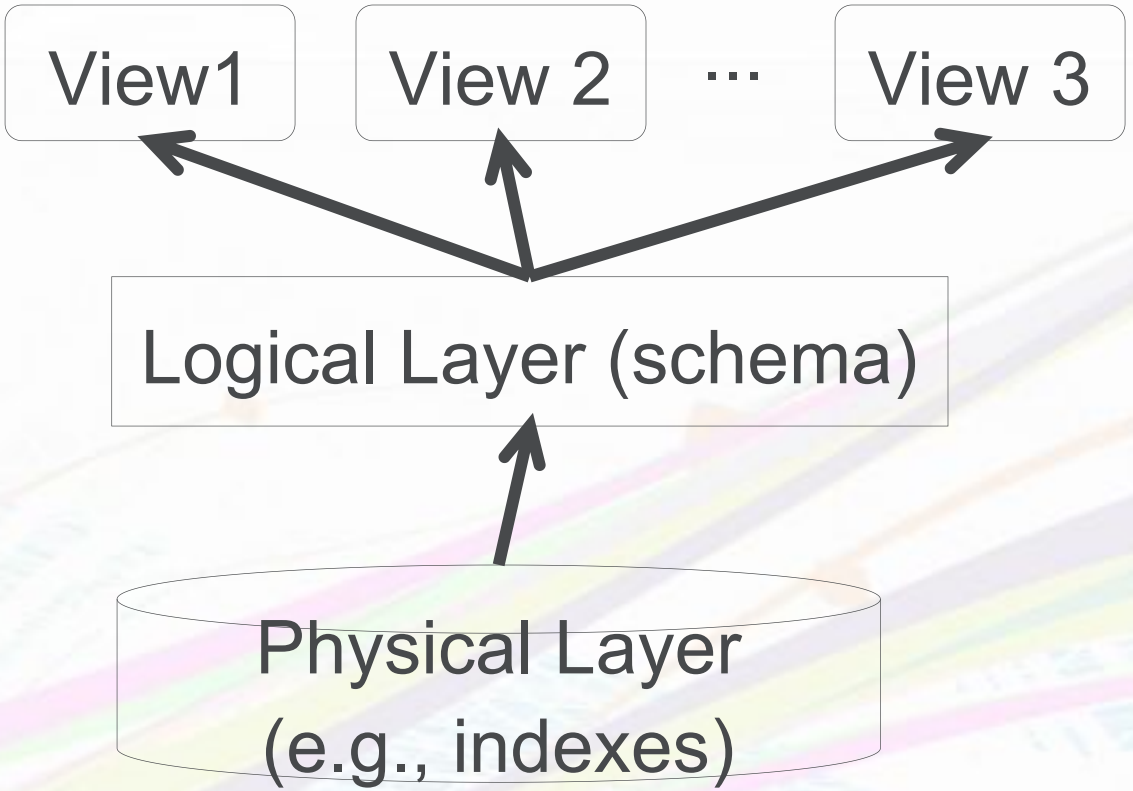


Database Abstraction Layers

Data Independence

Logical Data Independence

Physical Data Independence



Changes at one layer do not affect another layer!