Data Analytics Tools and Techniques

IOE 373, Fall 2023

Luis M Garcia-Guzman, Ph.D.

Instructors email: IOE373instructors@umich.edu

Some of my background..

Education Background:

- BS (ITESM-Mexico), MSE and PhD (U of M) --- Industrial Engineering

Work Experience:

- UM Lecturer, Undergrad Advisor, Researcher, Consultant – IOE
- Past: Product & Industrial Engineer ~ Duroplast (Injection Molding), Consultant in Quality Engineering for Chrysler, GM

Teaching Experience:

- IOE 201 – Economic Decision Making
- IOE 265 – Probability and Statistics
- IOE 366 – Linear Models
- IOE 373 – Data Processing
- IOE 465 – Design of Experiments
- IOE 466 – Statistical Process Control
- IOE 474 – Discrete Event Simulation
- Six Sigma Green Belt and Black Belt Courses

GSI/IA

Juhyun Lee - GSI
E-mail: juhyunl@umich.edu

Urvee Deo - IA
E-mail: urveedeo@umich.edu

- **Check class email, announcements, Piazza and Lecture Files**
- **Piazza: For HW/Lab assignments questions. We will Respond within 24 hrs during the weekdays.**
- **During office hours do not ask to debug your code, for coding issues not answered through Piazza, email instructors group.**

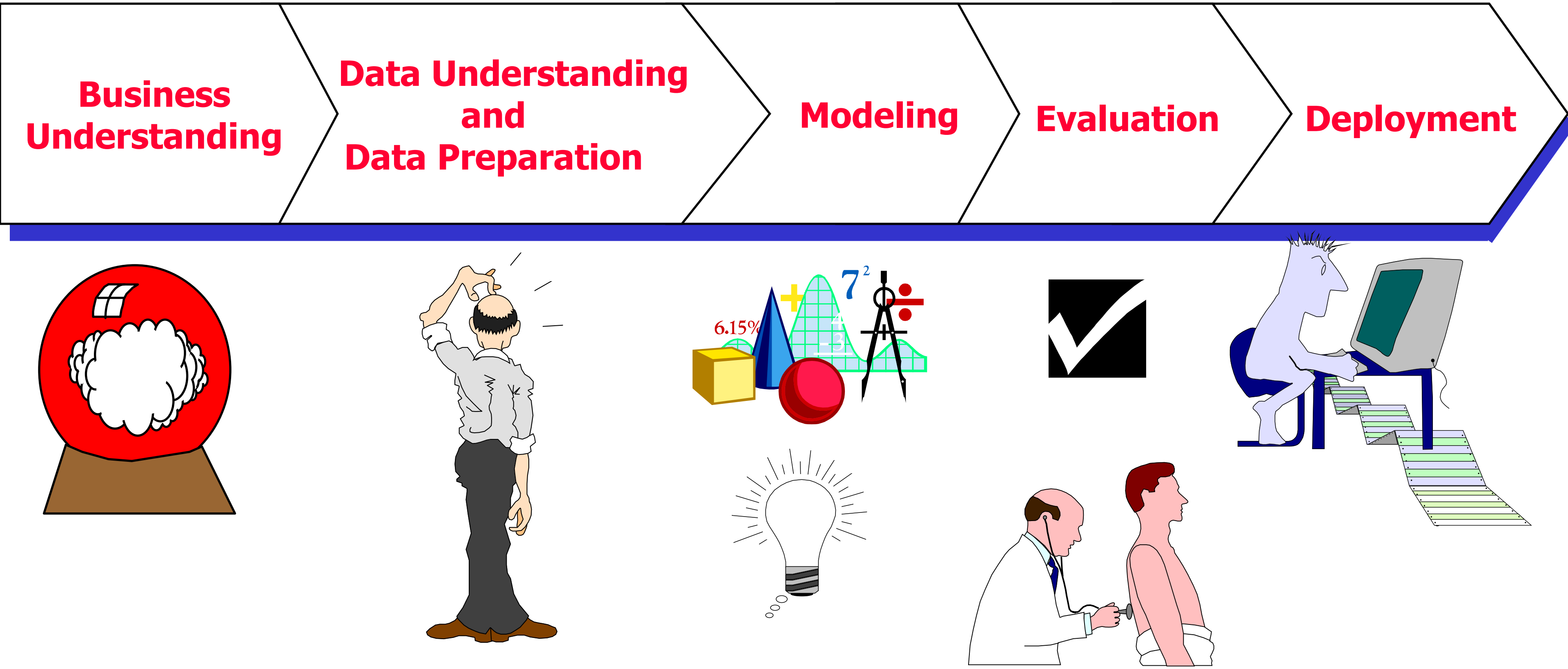
Reference Material/Books

- Beginning Database Design by Clare Churcher
  - https://search.lib.umich.edu/catalog/record/99187292424706381
- Excel 2019 Power Programming with VBA by Michael Alexander
  - https://search.lib.umich.edu/catalog/record/99187282400406381
- Python for Data Analysis by Wes McKinney
  - https://search.lib.umich.edu/catalog/record/99187294648606381
- Data Mining for Business Analytics: Concepts, Techniques and Applications in Python by Shmueli, et. al.
  - https://search.lib.umich.edu/catalog/record/99187538071606381

Course Topics

- Relational Database Design Concepts and Principles
- Access Database and Structured Query Language (SQL)
- Excel Visual Basic Programming
- Introduction to Python
- Introduction to Visualization and Analytics with Python

Data Analytics Process



Business Understanding

Data Understanding and Data Preparation

Modeling

Evaluation

Deployment

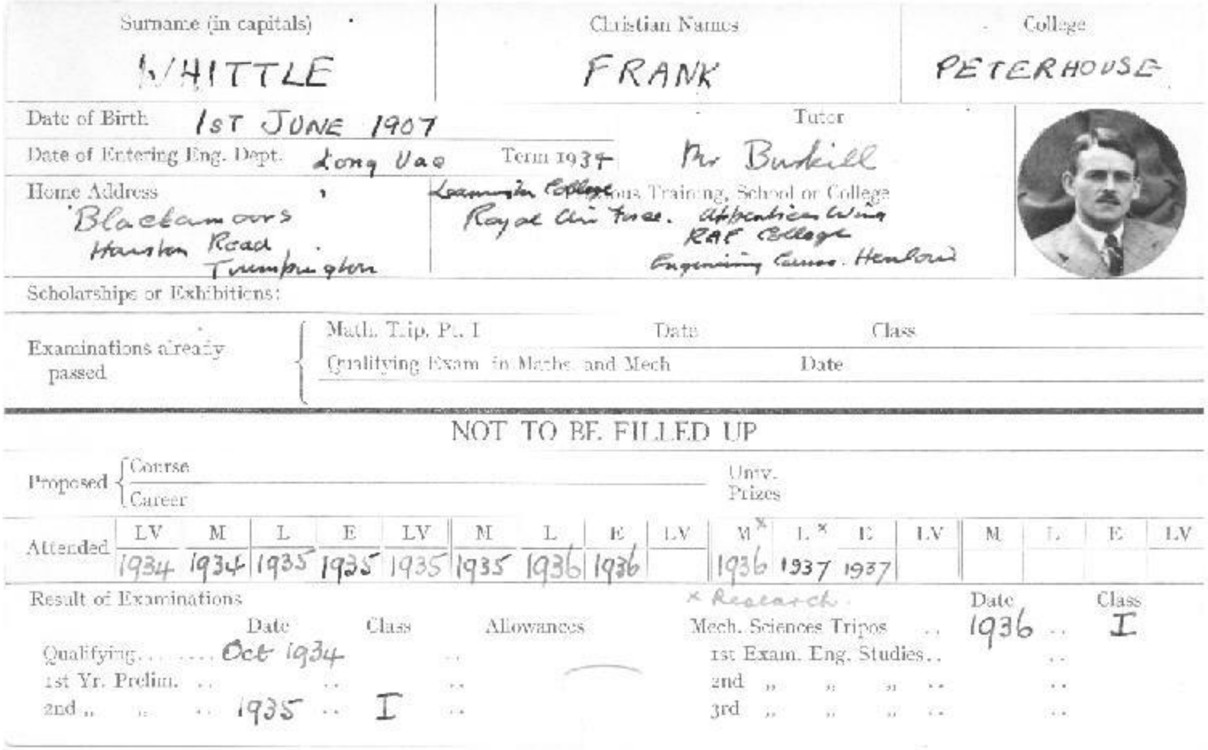
Introduction to Relational Databases

What is Data?

- Any collection of facts, numbers, texts that you are interested in
- Raw Data vs. Processed Data
- Questions:
  - What data should we keep?
  - How do we organize it?
  - In what ways do we use or display it
- These days a new term/concept: Big Data, Analytics…
- https://www.sas.com/en_us/insights/analytics/what-is-analytics.html

The Old Way

- Here’s an English university record from 1936:



Surname (in capitals) WHITTLE

Christian Names FRANK

College PETERHOUSE

Date of Birth 1st JUNE 1907

Date of Entering Eng. Dept. Long Vac Term 1934

Tutor Mr Burkill

Home Address 'Blackamoors Hauxton Road Trumpington

Leamington College Previous Training, School or College
Royal Air Force. Apprentice Wing
RAF College
Engineering Course. Henlow

Scholarships or Exhibitions:

Examinations already passed
Math. Trip. Pt. I Date Class
Qualifying Exam in Maths. and Mech Date

NOT TO BE FILLED UP

Proposed Course
Career

Univ. Prizes

Attended	LV	M	L	E	LV	M	L	E	LV	M×	L×	E	LV	M	L	E	LV
	1934	1934	1935	1935	1935	1935	1936	1936		1936	1937	1937					

Result of Examinations

× Research

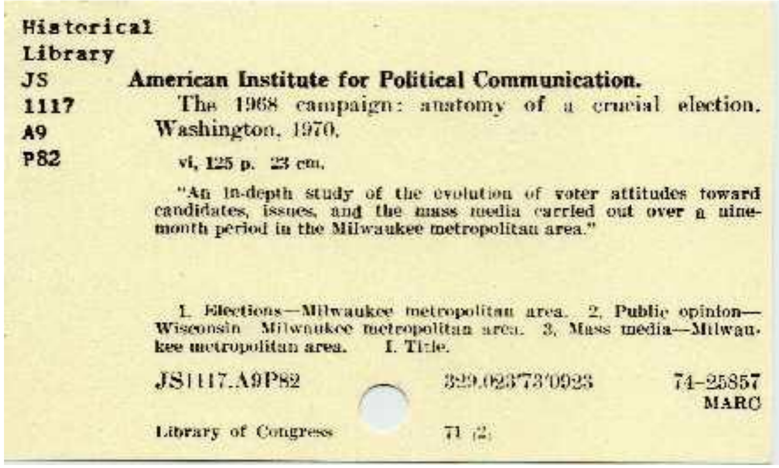
	Date	Class	Allowances
Qualifying	Oct 1934		
1st Yr. Prelim.			
2nd ,, ,,	1935	I	

	Date	Class
Mech. Sciences Tripos	1936	I
1st Exam. Eng. Studies		
2nd ,, ,, ,,		
3rd ,, ,, ,,		

Department of Engineering, Cambridge University
Photograph of student's record card

Some organizations had a lot of records to keep:

Library Card Catalogs



Historical
Library
JS 1117
A9
P82

American Institute for Political Communication.

The 1968 campaign: anatomy of a crucial election. Washington, 1970.

vi, 125 p. 23 cm.

"An in-depth study of the evolution of voter attitudes toward candidates, issues, and the mass media carried out over a nine-month period in the Milwaukee metropolitan area."

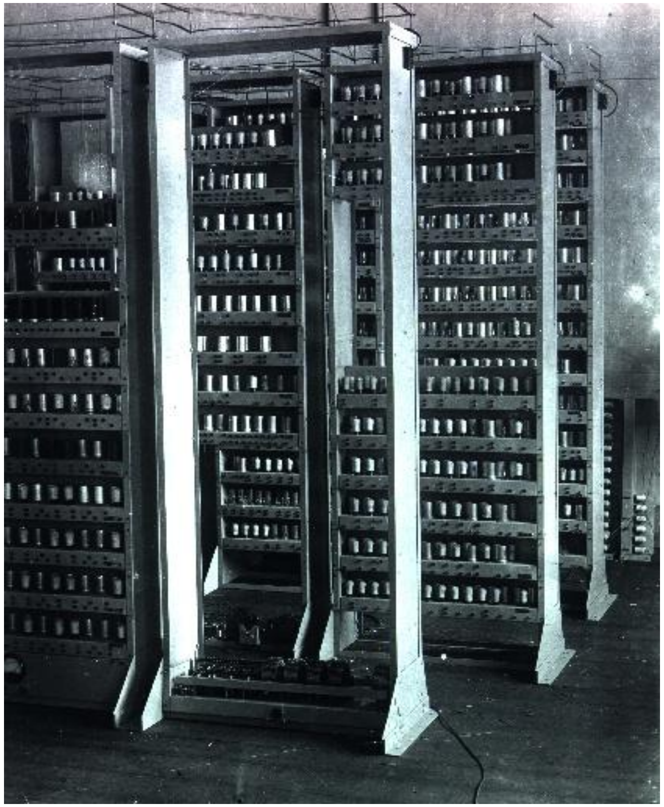
1. Elections—Milwaukee metropolitan area. 2. Public opinion—Wisconsin Milwaukee metropolitan area. 3. Mass media—Milwaukee metropolitan area. I. Title.

JS1117.A9P82 329.02373'0923 74-25857
MARC

Library of Congress 71 (2)

Enter the Computer

- Early electronic computers were used mainly for calculations, but this would change as the cost of storage media declined.



Stick that in your backpack!

- ***UNIVAC: a device which contained 20,000 vacuum tubes, occupied 1,500 square feet and weighed 40 tons***

Record Retrieval

- In the paper-record days, retrieving information was a physical process. A clerk had to:
  - Write down the information that was being requested, or have a client (colleague, boss, customer) fill out a form requesting the information.
  - Go to the appropriate building, room, and file cabinet.
  - Find the record in the file cabinet.
  - Find any associated data needed to complete the request, which might be in a different cabinet in a different room or building.
  - Take the records retrieved back to the office, and prepare the data for use by the client.
  - Return the records to their proper locations.
- If there was a backlog of requests, getting information from such a system could take days, weeks, or even months.

Compare to Today

- Suppose that a friend recommends a book to me: “Harry Potter and the Cursed Child” by JK Rowling.
- I can go to Amazon and find out all sorts of information about it: number of pages, cost, publisher, year published, etc.
- I can also find out related information:
  - Other books by this author.
  - What other readers say about this book.
  - What other books these readers recommend.

Or Movies

- I can search on my phone/laptop/smart TV on a number of different apps/websites and find out just about anything I want about any movie, actor, director, or movie theater.
- All within seconds! This sort of access to information just was not available 25 years ago.
- So what happened?

Hardware

- Many of these advances were made possible by improved and less-expensive hardware.
- Moore's Law: In 1965, Intel co-founder Gordon Moore noted that the number of transistors in an integrated circuit was doubling approximately every two years. This became known as Moore's Law, which has been generalized to refer to memory capacity and processor speed as well. The generalized form is that the technology of computing doubles in capacity every 18 months to 2 years. This has proven true for 50 years, and seems to be continuing into the future.

It wasn’t just hardware

- The earliest digital computers were developed in the 1940’s.
- It wasn’t until the 1960’s that computers began to be used in significant ways for record keeping—bank accounts, airline reservations, FBI files, and such.
- It wasn’t until the 1970’s that this recordkeeping was done in an organized fashion.
- It wasn’t until about 1995 that the World Wide Web went mainstream and brought widespread access to data to the masses.

Software

- Nevertheless, cheaper and faster computers were not the only reason for the revolution in data processing.
- Early attempts at developing methods for storing data encountered serious difficulties which limited their usefulness.
- You can read about these in the reading: DatabasesDemystifiedChapter1.pdf (available in Canvas).

Blue and Big Blue

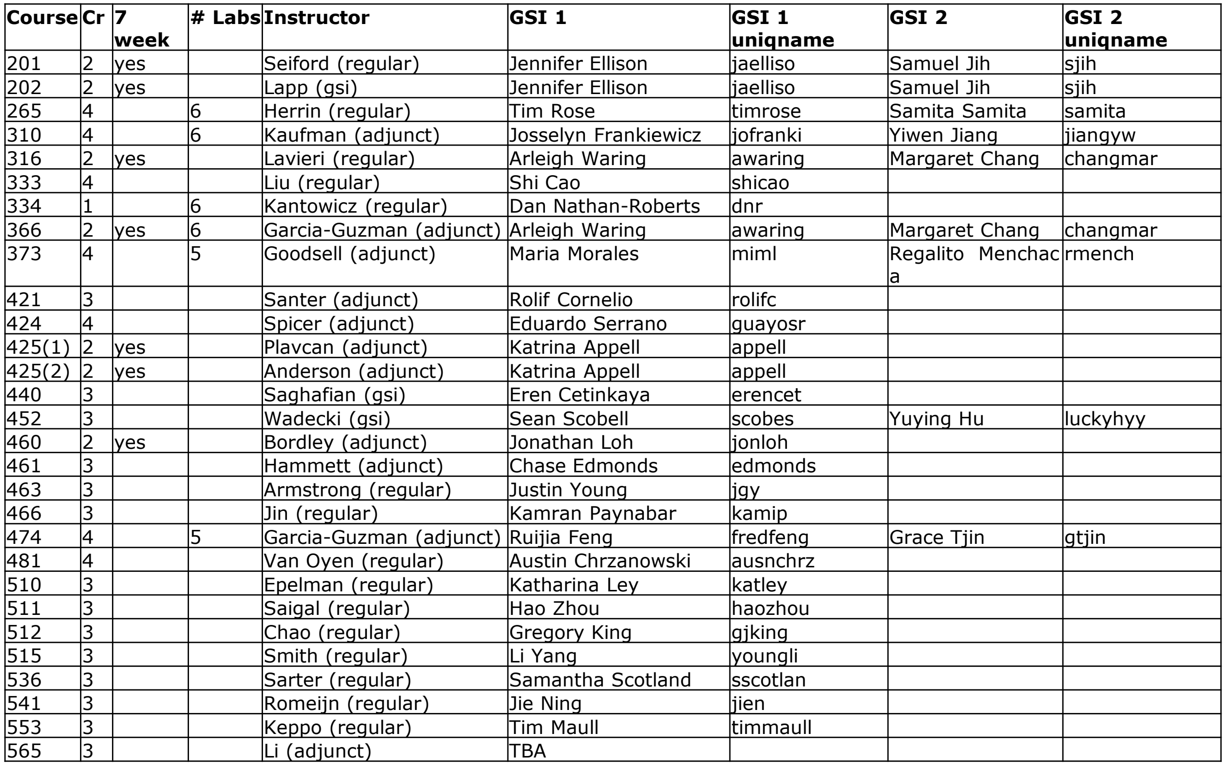
- In 1965, a young man named E.F. Codd earned his PhD here at the University of Michigan.
- Five years later he was working for IBM when he wrote a paper which is considered to be the origin of the relational database.
- The relational model quickly became the most successful way to store and access data on digital computers.

Some Definitions

- Database: A collection of interrelated data items that are managed as a single unit.
- Database Management System: The software application that organizes and retrieves data. Common DBMS’s are Oracle, Microsoft SQL Server, DB2, MySQL, and Access.
- Relational Database: A database consisting of tables which follow the rules specified by E.F. Codd.

Tables

- The simple model for what is called a “table” is an Excel spreadsheet:



Course	Cr	7 week	# Labs	Instructor	GSI 1	GSI 1 uniqname	GSI 2	GSI 2 uniqname
201	2	yes		Seiford (regular)	Jennifer Ellison	jaelliso	Samuel Jih	sjih
202	2	yes		Lapp (gsi)	Jennifer Ellison	jaelliso	Samuel Jih	sjih
265	4		6	Herrin (regular)	Tim Rose	timrose	Samita Samita	samita
310	4		6	Kaufman (adjunct)	Josselyn Frankiewicz	jofranki	Yiwen Jiang	jiangyw
316	2	yes		Lavieri (regular)	Arleigh Waring	awaring	Margaret Chang	changmar
333	4			Liu (regular)	Shi Cao	shicao		
334	1		6	Kantowicz (regular)	Dan Nathan-Roberts	dnr		
366	2	yes	6	Garcia-Guzman (adjunct)	Arleigh Waring	awaring	Margaret Chang	changmar
373	4		5	Goodsell (adjunct)	Maria Morales	miml	Regalito Menchaca	rmench
421	3			Santer (adjunct)	Rolif Cornelio	rolifc		
424	4			Spicer (adjunct)	Eduardo Serrano	guayosr		
425(1)	2	yes		Plavcan (adjunct)	Katrina Appell	appell		
425(2)	2	yes		Anderson (adjunct)	Katrina Appell	appell		
440	3			Saghafian (gsi)	Eren Cetinkaya	erencet		
452	3			Wadecki (gsi)	Sean Scobell	scobes	Yuying Hu	luckyhyy
460	2	yes		Bordley (adjunct)	Jonathan Loh	jonloh		
461	3			Hammett (adjunct)	Chase Edmonds	edmonds		
463	3			Armstrong (regular)	Justin Young	jgy		
466	3			Jin (regular)	Kamran Paynabar	kamip		
474	4		5	Garcia-Guzman (adjunct)	Ruijia Feng	fredfeng	Grace Tjin	gtjin
481	4			Van Oyen (regular)	Austin Chrzanowski	ausnchrz		
510	3			Epelman (regular)	Katharina Ley	katley		
511	3			Saigal (regular)	Hao Zhou	haozhou		
512	3			Chao (regular)	Gregory King	gjking		
515	3			Smith (regular)	Li Yang	youngli		
536	3			Sarter (regular)	Samantha Scotland	sscotlan		
541	3			Romeijn (regular)	Jie Ning	jien		
553	3			Keppo (regular)	Tim Maull	timmaull		
565	3			Li (adjunct)	TBA			

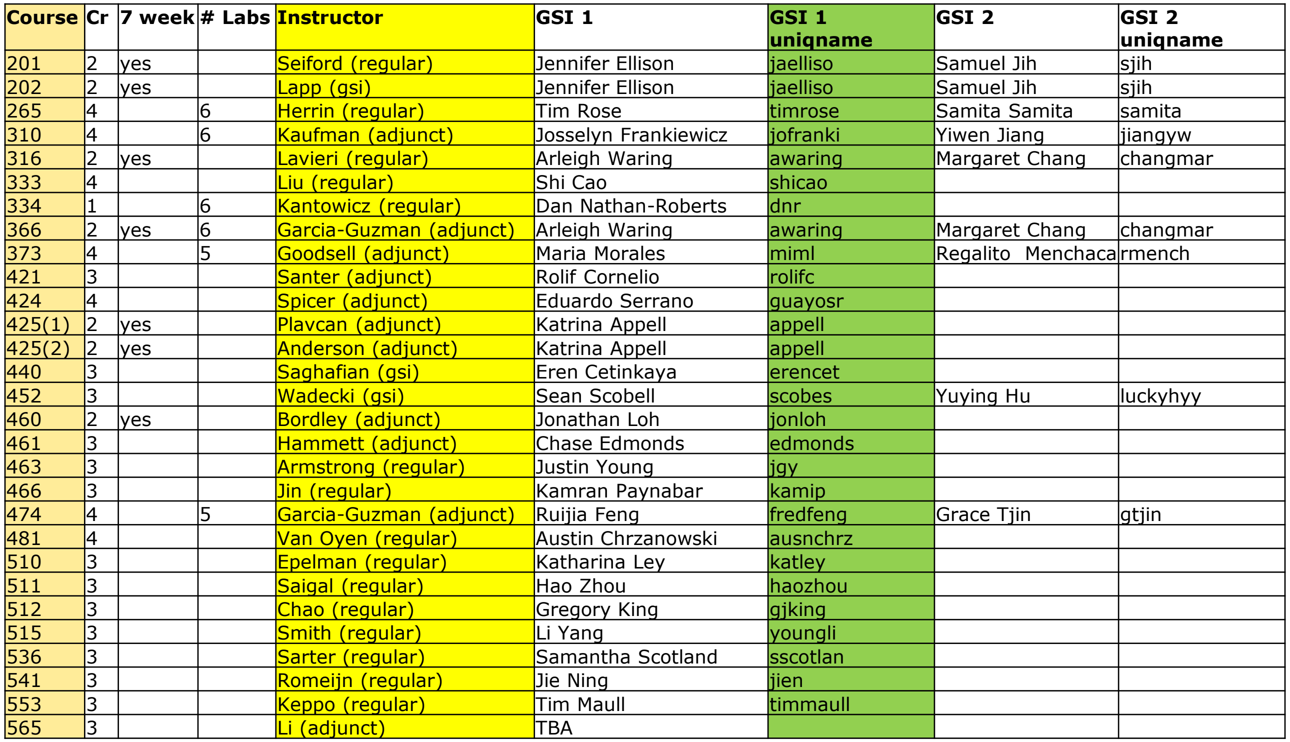
- A rectangular grid of data, with each column representing a property belonging to each row.

Table Definitions

- The spreadsheet on the previous slide is not actually a proper table for a database.
- By next week, you’ll know what I mean by that.
- For now, let’s start with some definitions. We’ll define “table” later; let’s look at the parts of a table first:

Field

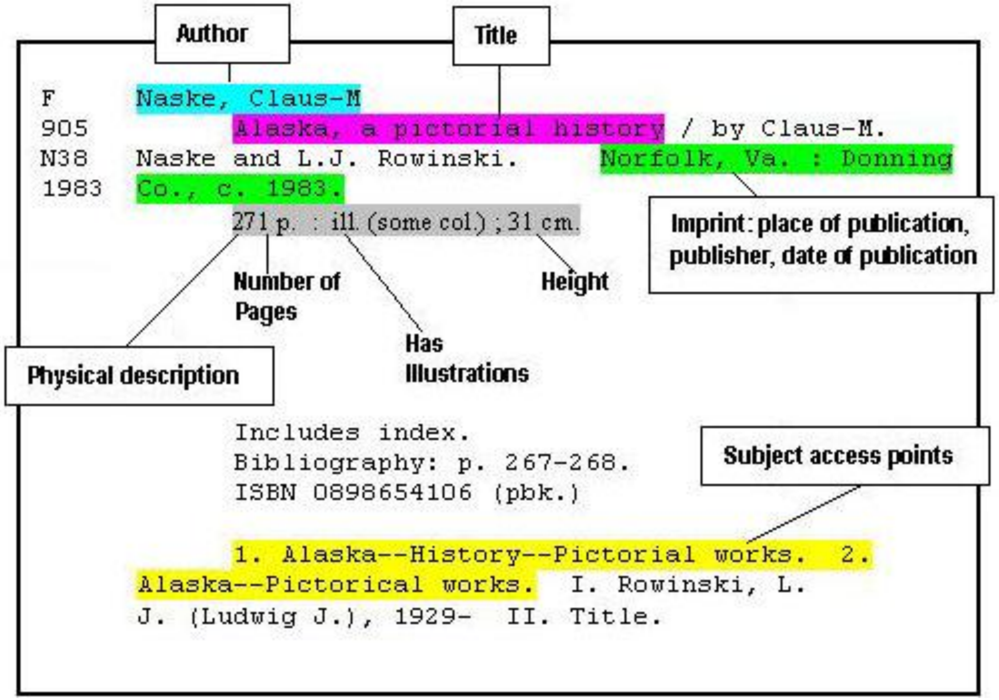
- Field: One of the vertical columns in a table. The terms “field” and “column” are used interchangeably when talking about tables; in more theoretical discussions, the term “attribute” is used.



Course	Cr	7 week	# Labs	Instructor	GSI 1	GSI 1 uniqname	GSI 2	GSI 2 uniqname
201	2	yes		Seiford (regular)	Jennifer Ellison	jaelliso	Samuel Jih	sjih
202	2	yes		Lapp (gsi)	Jennifer Ellison	jaelliso	Samuel Jih	sjih
265	4		6	Herrin (regular)	Tim Rose	timrose	Samita Samita	samita
310	4		6	Kaufman (adjunct)	Josselyn Frankiewicz	jofranki	Yiwen Jiang	jiangyw
316	2	yes		Lavieri (regular)	Arleigh Waring	awaring	Margaret Chang	changmar
333	4			Liu (regular)	Shi Cao	shicao		
334	1		6	Kantowicz (regular)	Dan Nathan-Roberts	dnr		
366	2	yes	6	Garcia-Guzman (adjunct)	Arleigh Waring	awaring	Margaret Chang	changmar
373	4		5	Goodsell (adjunct)	Maria Morales	miml	Regalito Menchaca	rmench
421	3			Santer (adjunct)	Rolif Cornelio	rolifc		
424	4			Spicer (adjunct)	Eduardo Serrano	guayosr		
425(1)	2	yes		Plavcan (adjunct)	Katrina Appell	appell		
425(2)	2	yes		Anderson (adjunct)	Katrina Appell	appell		
440	3			Saghafian (gsi)	Eren Cetinkaya	erencet		
452	3			Wadecki (gsi)	Sean Scobell	scobes	Yuying Hu	luckyhyy
460	2	yes		Bordley (adjunct)	Jonathan Loh	jonloh		
461	3			Hammett (adjunct)	Chase Edmonds	edmonds		
463	3			Armstrong (regular)	Justin Young	jgy		
466	3			Jin (regular)	Kamran Paynabar	kamip		
474	4		5	Garcia-Guzman (adjunct)	Ruijia Feng	fredfeng	Grace Tjin	gtjin
481	4			Van Oyen (regular)	Austin Chrzanowski	ausnchrz		
510	3			Epelman (regular)	Katharina Ley	katley		
511	3			Saigal (regular)	Hao Zhou	haozhou		
512	3			Chao (regular)	Gregory King	gjking		
515	3			Smith (regular)	Li Yang	youngli		
536	3			Sarter (regular)	Samantha Scotland	sscotlan		
541	3			Romeijn (regular)	Jie Ning	jien		
553	3			Keppo (regular)	Tim Maull	timmaull		
565	3			Li (adjunct)	TBA			

Field

- Note that the term “Field” comes from the blanks that needed to be filled in on paper forms: “Fill in all of the fields on the form.”



F Naske, Claus-M
905 Alaska, a pictorial history / by Claus-M.
N38 Naske and L.J. Rowinski. Norfolk, Va. : Donning
1983 Co., c. 1983.
271 p. : ill. (some col.) ; 31 cm.

Includes index.
Bibliography: p. 267-268.
ISBN 0898654106 (pbk.)

1. Alaska--History--Pictorial works. 2.
Alaska--Pictorical works. I. Rowinski, L.
J. (Ludwig J.), 1929- II. Title.

Author
Title
Imprint: place of publication, publisher, date of publication
Number of Pages
Height
Has Illustrations
Physical description
Subject access points

Record

- A “record” is like an individual form in a file cabinet or card in a card catalog.
- It represents one particular item of a type: one student, one book, etc.
- A record has particular values for the various fields: the student’s name, uniqname, ID, phone number, etc.
- In a spreadsheet or table, records are represented by horizontal rows.
- Therefore, the terms “record” and “row” are used interchangeably in discussing databases.

Records

Course	Cr	7 week	# Labs	Instructor	GSI 1	GSI 1 uniqname	GSI 2	GSI 2 uniqname
201	2	yes		Seiford (regular)	Jennifer Ellison	jaelliso	Samuel Jih	sjih
202	2	yes		Lapp (gsi)	Jennifer Ellison	jaelliso	Samuel Jih	sjih
265	4		6	Herrin (regular)	Tim Rose	timrose	Samita Samita	samita
310	4		6	Kaufman (adjunct)	Josselyn Frankiewicz	jofranki	Yiwen Jiang	jiangyw
316	2	yes		Lavieri (regular)	Arleigh Waring	awaring	Margaret Chang	changmar
333	4			Liu (regular)	Shi Cao	shicao		
334	1		6	Kantowicz (regular)	Dan Nathan-Roberts	dnr		
366	2	yes	6	Garcia-Guzman (adjunct)	Arleigh Waring	awaring	Margaret Chang	changmar
373	4		5	Goodsell (adjunct)	Maria Morales	miml	Regalito Menchaca	rmench
421	3			Santer (adjunct)	Rolif Cornelio	rolifc		
424	4			Spicer (adjunct)	Eduardo Serrano	guayosr		
425(1)	2	yes		Plavcan (adjunct)	Katrina Appell	appell		
425(2)	2	yes		Anderson (adjunct)	Katrina Appell	appell		
440	3			Saghafian (gsi)	Eren Cetinkaya	erencet		
452	3			Wadecki (gsi)	Sean Scobell	scobes	Yuying Hu	luckyhyy
460	2	yes		Bordley (adjunct)	Jonathan Loh	jonloh		
461	3			Hammett (adjunct)	Chase Edmonds	edmonds		
463	3			Armstrong (regular)	Justin Young	jgy		
466	3			Jin (regular)	Kamran Paynabar	kamip		
474	4		5	Garcia-Guzman (adjunct)	Ruijia Feng	fredfeng	Grace Tjin	gtjin
481	4			Van Oyen (regular)	Austin Chrzanowski	ausnchrz		
510	3			Epelman (regular)	Katharina Ley	katley		
511	3			Saigal (regular)	Hao Zhou	haozhou		
512	3			Chao (regular)	Gregory King	gjking		
515	3			Smith (regular)	Li Yang	youngli		
536	3			Sarter (regular)	Samantha Scotland	sscotlan		
541	3			Romeijn (regular)	Jie Ning	jien		
553	3			Keppo (regular)	Tim Maull	timmaull		
565	3			Li (adjunct)	TBA			

Summary of Definitions

- The vertical columns in a table are called “fields”;
- The horizontal rows are called “records”;
- Each record has values for each field—defining characteristics (attributes) which distinguish that record (student, book, whatever) from others in the table.

Introduction to Relational Databases

Theory and Practice

- We will begin by looking at the theory and the terminology of table design.
- After that, we will focus on the practical side—using Access to:
  - Create a new database
  - Design tables
  - Create relationships

Modern Databases

- Definition from “Databases Demystified‟: a database is a collection of interrelated data items that are managed as a single unit.
- This definition is deliberately vague, allowing it to cover most of the different types of databases that have been used over the past five decades or so.
- For a relational database, the definition can be more focused, at least on the logical level:
- Database: A collection of tables, the relationships between them, and auxiliary items such as views and stored procedures.

DBMS

- A database is managed, strangely enough, by something called a “database management system” (DBMS).
- Popular DBMS’s include Oracle, MySql, DB2, and Microsoft’s SQL Server (for large-scale databases) and Access (for smaller databases).
- Large-scale DBMS’s like Oracle and SQL Server typically run on specialized computers called servers, which provide data for many computers (clients) over networks. They typically store their data in many files, frequently spread across many hard drives, and even many different servers.

Access

- While Access can be used on networks, it is more of a “personal” DBMS, running on the user’s computer instead of a separate server. It stores everything in a single file (*.mdb for Access 2003 and earlier; *.accdb for Access 2007 and later).
- The single-file feature of Access is why we will use it in this class. It makes it easy for me to share entire databases with you, and for you to turn in databases for assignments and projects.

Microsoft Access

- Access is not the best or most powerful DBMS; it is just the most convenient for use in this class (and in a lot of companies).
- Nevertheless, it is good enough and powerful enough that it serves very nicely as a training database: most of what you need to know about databases you can learn using Access.
- You'll start learning how to use Access in the next lecture and in lab.

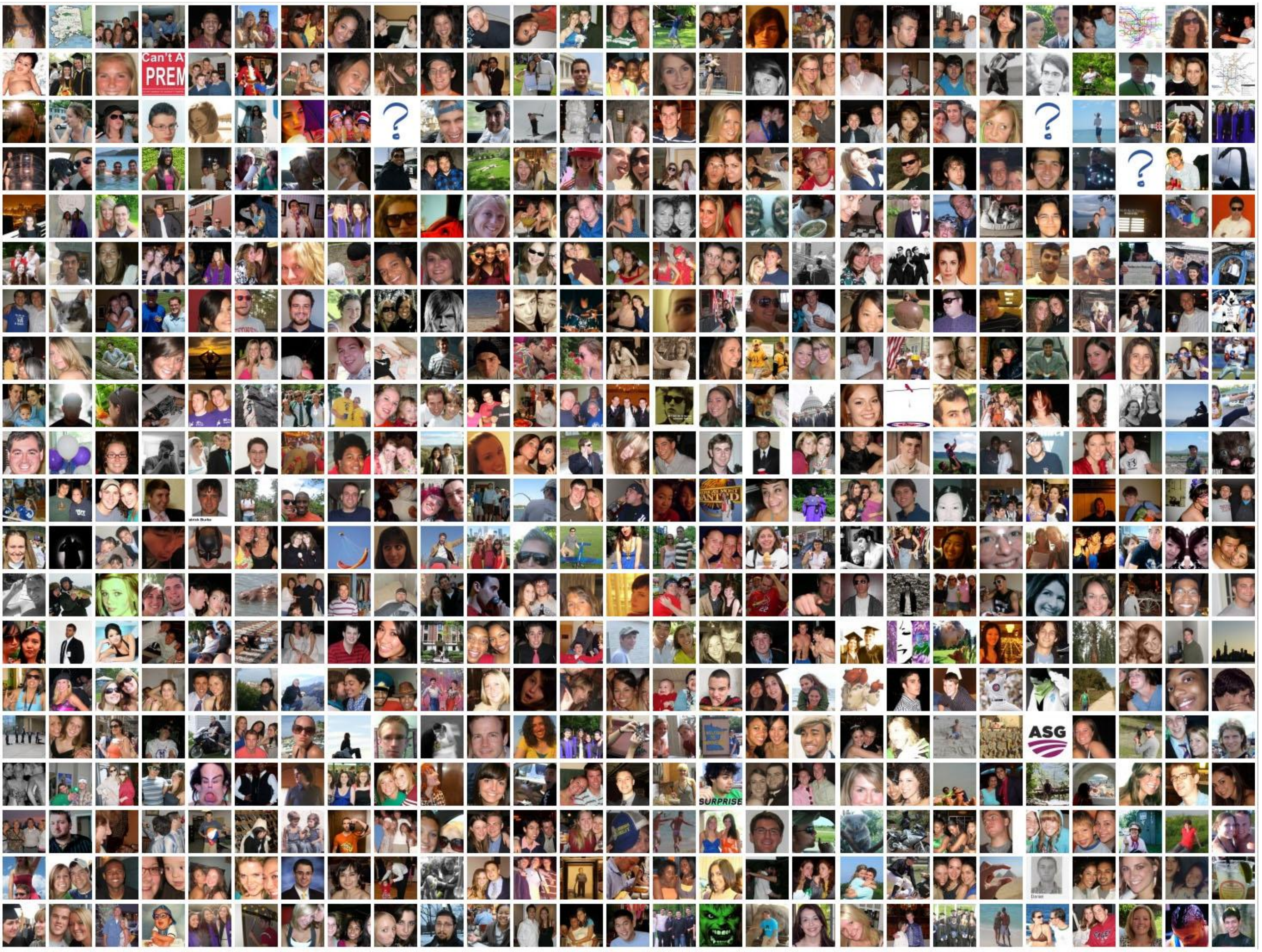
Key Point!

- Good relational database design is about optimizing how the data is STORED
- Most “tables” you have seen—in books, in lectures, on the web—were probably optimized for display, not for storage.
- Relational database tables are designed for consistency and to reduce redundancy. They are not designed for appearance.

Why “relational?”

- Relationships are not what gave the relational database its name.
- The term “relational” comes from the mathematical concept of “relation,” which refers to a set of ordered pairs (or triplets, etc.; the generic term is “tuple”) of items. A mathematical function is a special type of relation.

Example – Social Media (e.g Instagram)



How to Organize Data

- Let’s look at an extremely simplified version of a social media app where you register and then want to “follow” other people:

ID	Name	Gender	Email	Origin
1	Philip J. Fry	M	pfry@futurama.com	New York City, New York, U.S
2	Tanya Leela	F	tleela@futurama.com	New New York City, New New York
3	Benjamin Rodriguez	M	bender@futurama.com	Tijuana, Baja California, Mexico
4	Professor Farnsworth	M	farnsworth@futurama.com	New New York City, NewNew York
5	Doctor John Zoidberg	M	zoidberg@futurama.com	Miami, FL
6	Amy Wong	F	awong@futurama.com	Dallas, TX

How to Represent who you follow?

- If we are using Excel, one may easily come up with this

ID	Name	Gender	Email	Origin	Follow1	Follow2	Follow3	Follow4
1	Philip J. Fry	M	pfry@futurama.com	New York City, New York, U.S	Tanya Leela	Benjamin Rodriguez	Professor Farnsworth	Amy Wong
2	Tanya Leela	F	tleela@futurama.com	New New York City, New New York	Philip J. Fry	Amy Wong	Professor Farnsworth	Doctor John Zoidberg
3	Benjamin Rodriguez	M	bender@futurama.com	Tijuana, Baja California, Mexico	Professor Farnsworth	Doctor John Zoidberg	Philip J. Fry	
4	Professor Farnsworth	M	farnsworth@futurama.com	New New York City, NewNew York	Philip J. Fry	Tanya Leela	Bendjamin Rodriguez	Doctor John Zoidberg
5	Doctor John Zoidberg	M	zoidberg@futurama.com	Miami, FL	Amy Wong	Tanya Leela	Benjamin Rodriguez	Professor Farnsworth
6	Amy Wong	F	awong@futurama.com	Dallas, TX	Philip J. Fry	Tanya Leela	Doctor John Zoidberg	

Another Version….

ID	Name	Gender	Email	Origin	Follow 1	Follow 2	Follow 3	Follow 4
1	Philip J. Fry	M	pfry@futurama.com	New York City, New York, U.S	2	3	4	6
2	Tanya Leela	F	tleela@futurama.com	New New York City, New New York	1	6	4	5
3	Benjamin Rodriguez	M	bender@futurama.com	Tijuana, Baja California, Mexico	4	5	1	
4	Professor Farnsworth	M	farnsworth@futurama.com	New New York City, NewNew York	1	2	3	5
5	Doctor John Zoidberg	M	zoidberg@futurama.com	Miami, FL	6	2	3	4
6	Amy Wong	F	awong@futurama.com	Dallas, TX	1	2	5	

What If More People Are Joining

- You’d need to keep increasing the number of columns
- A lot of redundant columns (some people follow few people)

Improvement

- Use two tables named “Person” and “Friendship”:

Person

ID	Name	Gender	Email	Origin
1	Philip J. Fry	M	pfry@futurama.com	New York City, New York, U.S
2	Tanya Leela	F	tleela@futurama.com	New New York City, New New York
3	Benjamin Rodriguez	M	bender@futurama.com	Tijuana, Baja California, Mexico
4	Professor Farnsworth	M	farnsworth@futurama.com	New New York City, NewNew York
5	Doctor John Zoidberg	M	zoidberg@futurama.com	Miami, FL
6	Amy Wong	F	awong@futurama.com	Dallas, TX

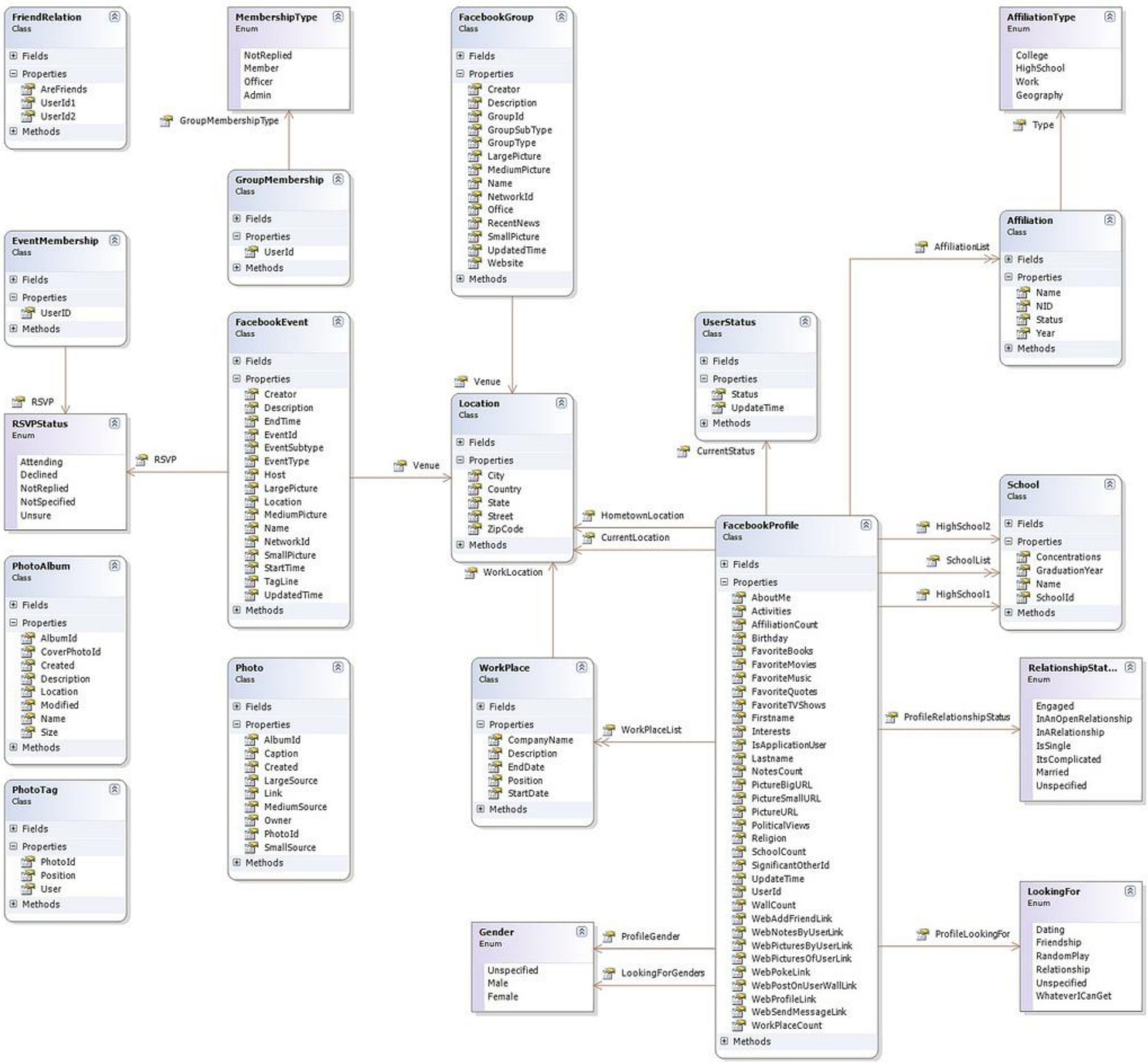
“Friendship”

ID1	ID2
1	2
1	3
1	4
1	6
2	4
2	5
2	6
3	4
3	5
4	5
5	6

Advantages

- Can be easily scaled-up (increase size)
- No redundancy

An Overall Picture



By Anatoly Lubarsky: http://blogs.x2line.com/al/archive/2007/06/02/3124.aspx

Design and Maintain a Database

- We are going to formalize the “tricks” we used in the example as database design techniques

Theory and Practice

- Theory Entitiy-Relationship Diagram
  - Entitiy-Relationship Diagram (ERD)
  - Normal Forms (NF): 1NF, 2NF, 3NF, BCNF
- Practice
  - Creating and Managing a Database
  - Designing Data-driven Application

Dealing With Data

- What Data We Need
- How to Organize These Data
- How to Store These Data
- How to Retrieve Data
- How to Analyze and Display Data

ERD - Relational Database Design Tool

- **Entity-Relationship Diagram** is a powerful tool to help you understand database concepts and designing a proper database

Entity

- An entity is something about which we store data
- For example, a customer is an entity
- Entities are not necessarily tangible. For example, a doctor’s appointment can be an entity

Entity

- In a properly designed relational database, each relation (table) represents a single “entity”.
- An entity is sort of a generic noun. For example, the concept of Customer is an entity, but one particular customer is not an entity.
- In object-oriented programming (OOP), an entity is typically represented by something called a “class.” An individual instance of that class (a particular customer, for example), is called an “object.”

Attribute

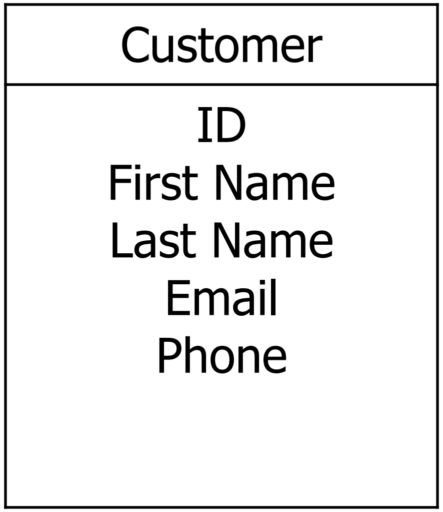
- Attributes are what describe an entity
- For example, a customer entity is usually described by a customer number, first name, last name, email, phone number, etc.
- When we represent entities in a database, we actually store only the attributes

Terminology Comparison

Object-Oriented Program	**Database Design**	**A.K.A**
Class	Entity	Table
Object	Instance	Row (Record)
Property	Attribute	Column (Field)

Diagram to Represent Entity

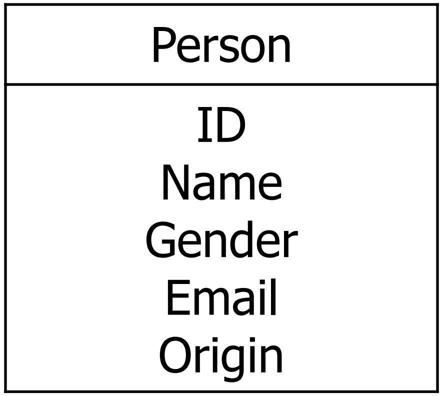
- We use the following diagram to represent an entity:



Customer

ID
First Name
Last Name
Email
Phone

Entities in “Instagram” Example

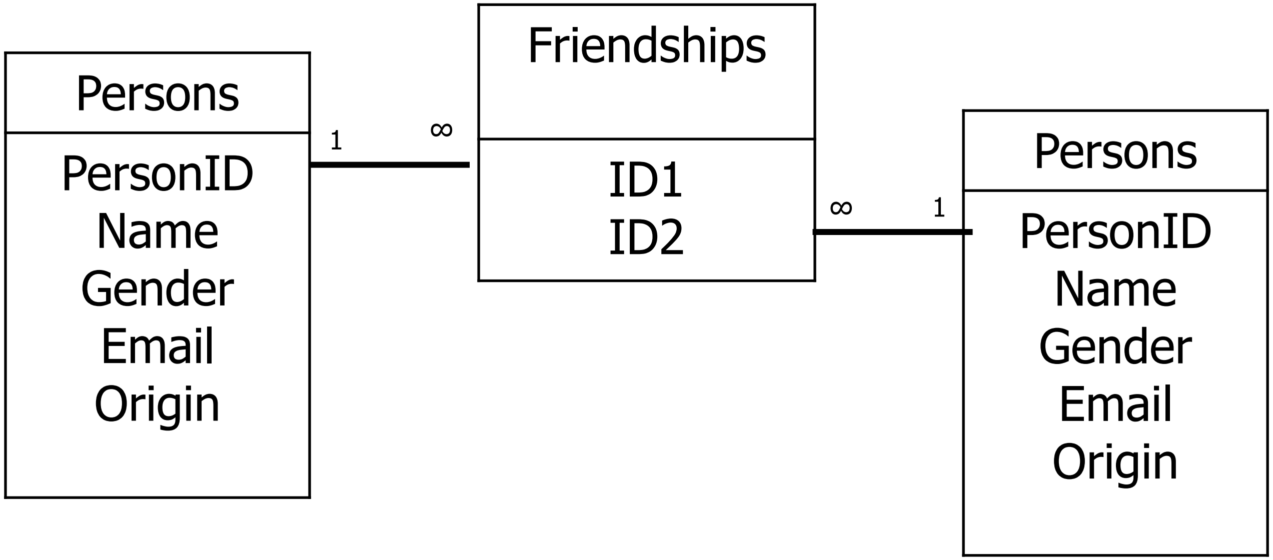


Person

ID
Name
Gender
Email
Origin

Relationship of Entities

- We can draw a line between attributes in different entities to represent the connection they have



Persons

PersonID
Name
Gender
Email
Origin

1

∞

Friendships

ID1
ID2

∞

1

Persons

PersonID
Name
Gender
Email
Origin