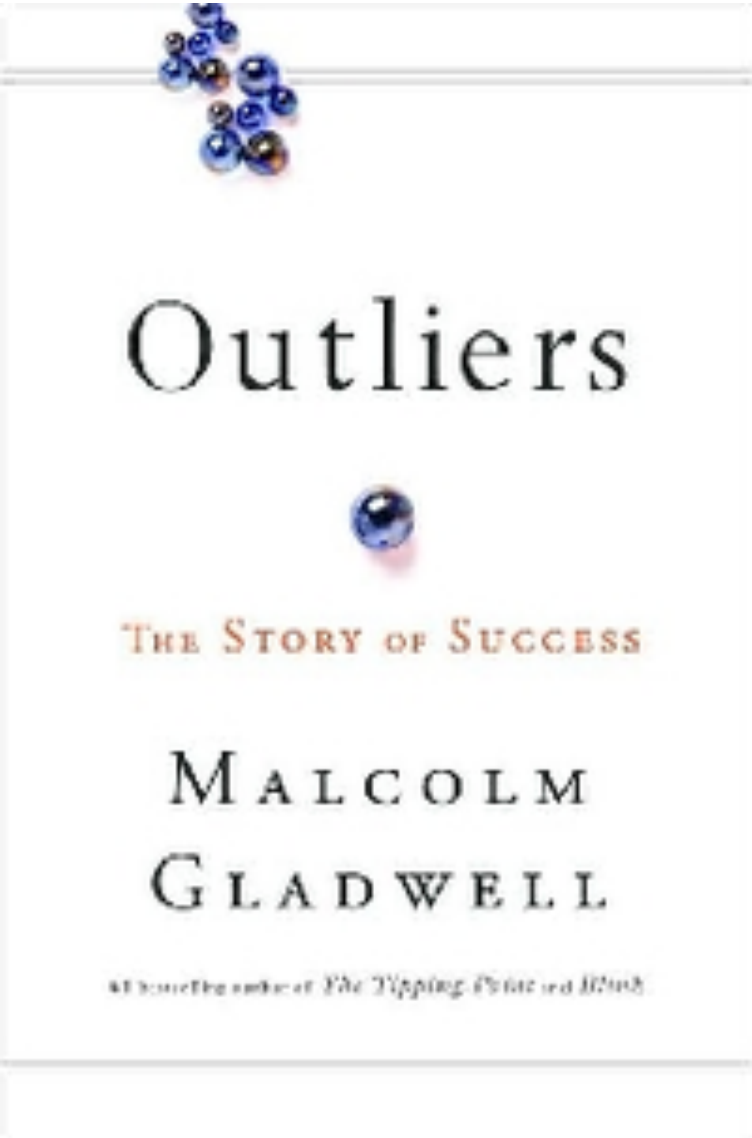
Technology Fundamentals for Analytics

Jason Kuruzovich

What is the Significance of 10,000 Hours???

Outliers



Throughout the publication, Gladwell repeatedly mentions the "10,000-Hour Rule", claiming that the key to success in any field is, to a large extent, a matter of practicing a specific task for a total of around 10,000 hours.

Article

How to hire data scientists and get hired as one

- SQL,
- Statistics,
- Predictive modeling and
- Programming (probably Python)

Course Wiki [Canvas -> Pages]

- Tons of things to learn out there and many more free resources than ever before in the history of the world
- Help me by adding good resources that you find.

Last Time

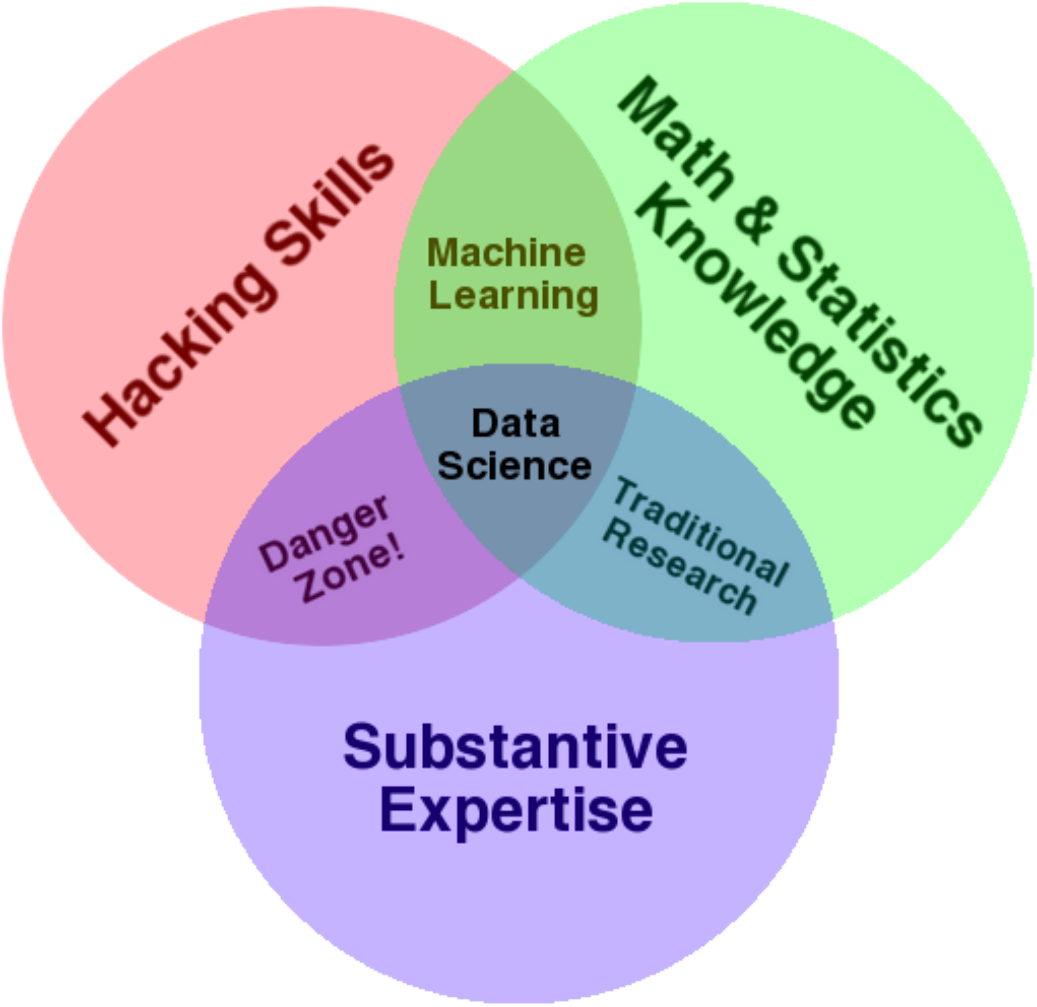
- Overview of Semester
  - What does it mean to be a data scientist?
  - What do we mean by analytics?
  - What we we mean by big data?

QUIZ: WHAT ARE THE 3 PRIMARY AREAS OF EXPERTISE FOR A DATA SCIENTIST?

Key Tools of the Data Scientist

- Data Munging - parsing, scraping, and formatting data
- Math and Statistics - traditional analysis you're used to thinking about
- Business Expertise – Knowledge of the business domain

Data Science Venn Diagram



Source:
http://drewconway.com/zia/2013/3/26/the-data-science-venn-diagram

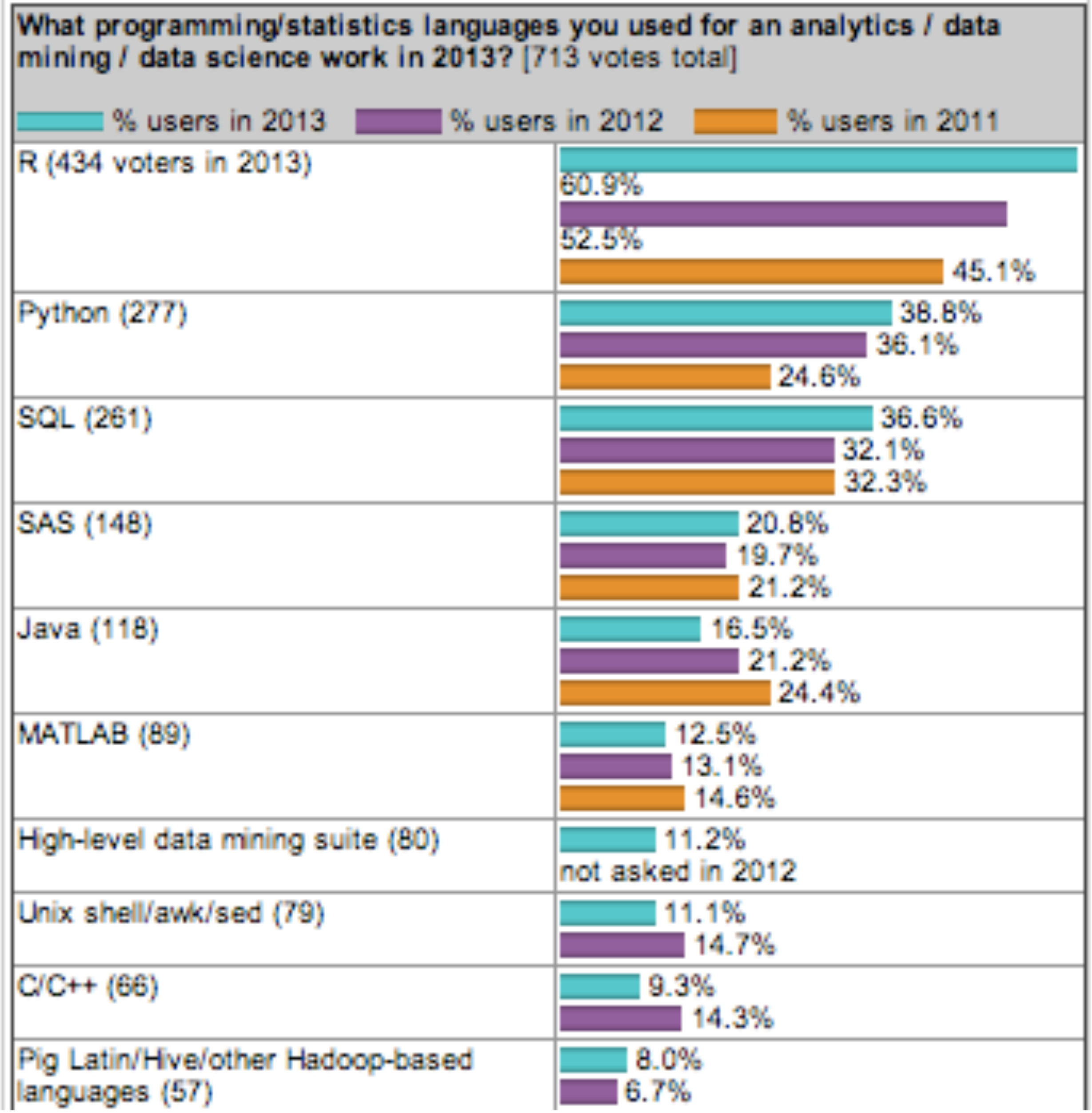
Background

What is R?

R is a system for statistical computation and graphics. It consists of a language plus a run-time environment with graphics, a debugger, access to certain system functions, and the ability to run programs stored in script files.

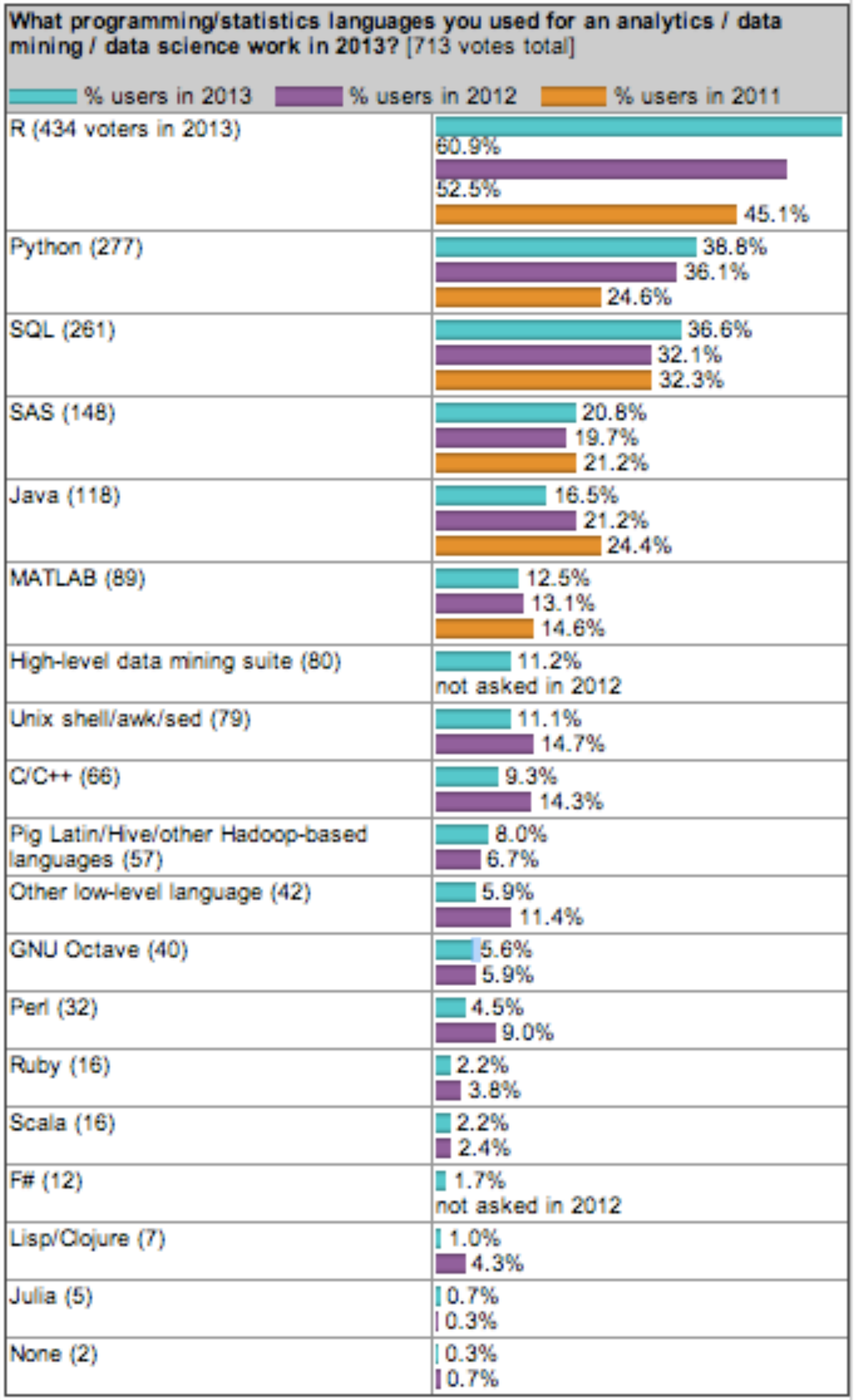
What is CRAN?

The “Comprehensive R Archive Network” (CRAN) is a collection of sites which carry identical material, consisting of the R distribution(s), the contributed extensions, documentation for R, and binaries.



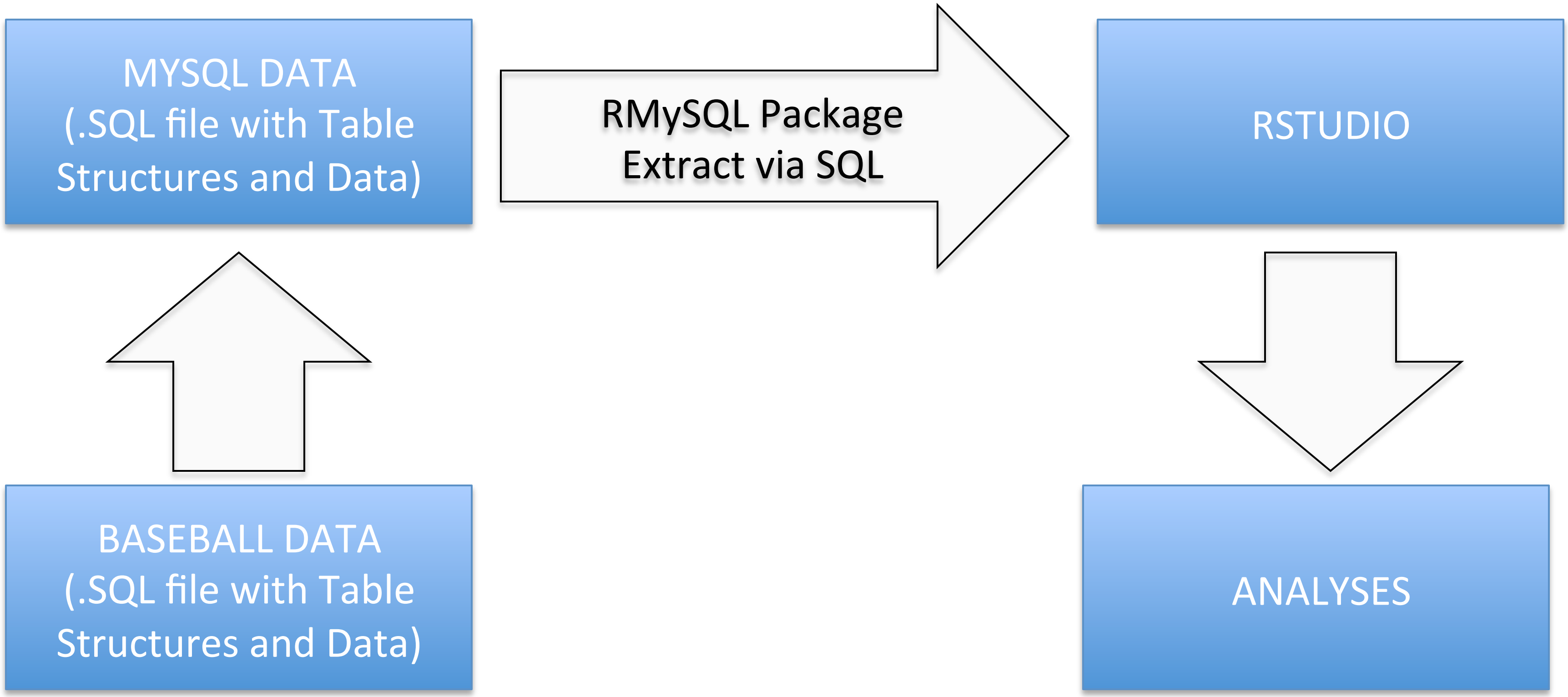
R is Top Language for Data Mining / Data Science Work

R is Top Language for Data Mining / Data Science Work



Lab, What we are Doing

- Lab – Foundations and Key Ideas



MYSQL DATA
(.SQL file with Table Structures and Data)

RMySQL Package
Extract via SQL

RSTUDIO

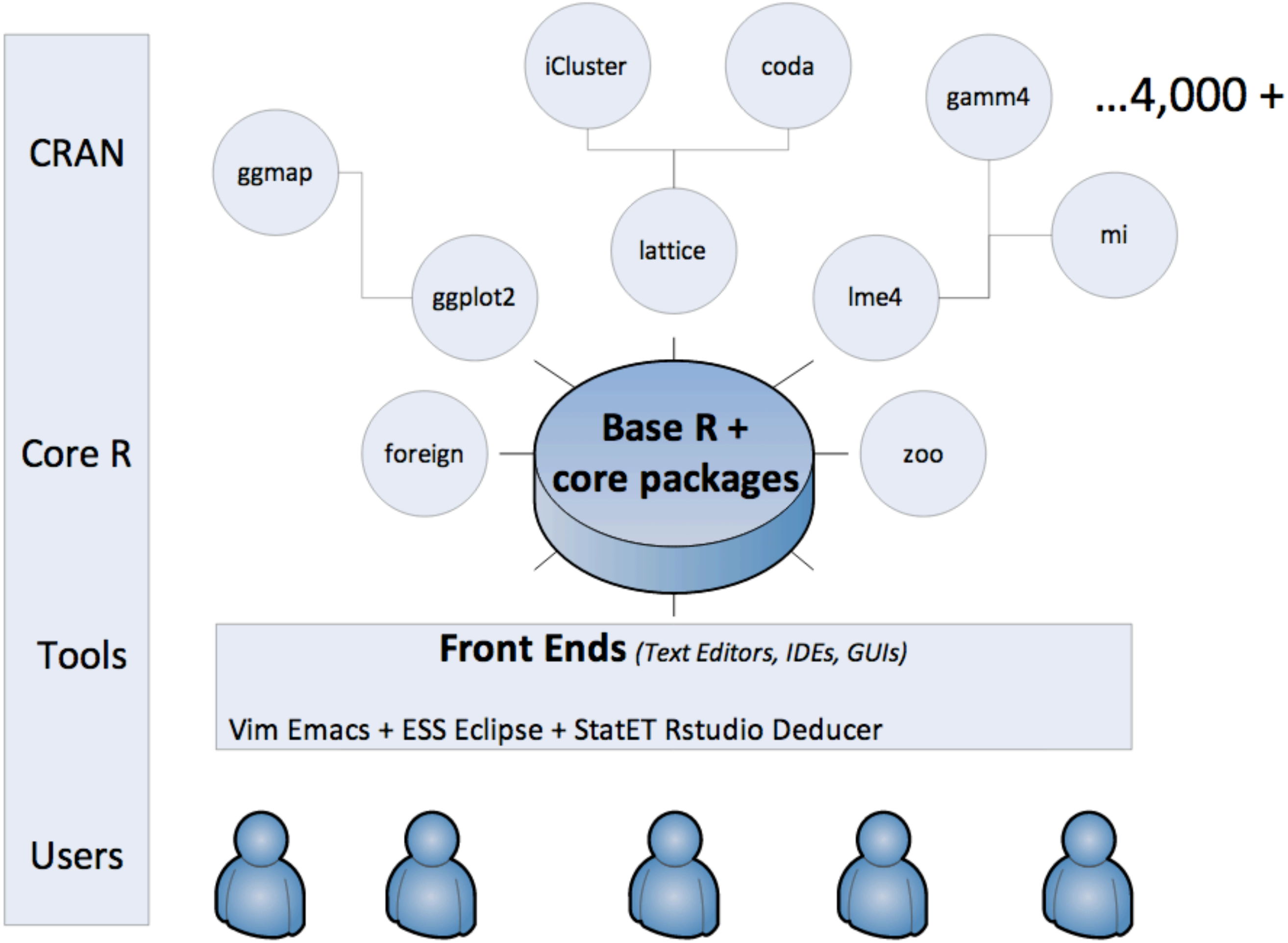
BASEBALL DATA
(.SQL file with Table Structures and Data)

ANALYSES

R and Packages

- Packages are collections of **R** functions, data, and compiled code in a well-defined format.
- **R** comes with a standard set of packages.
- Others are available for download and installation.
- External packages only have to be installed once, but they have to be loaded each time they are used.
- You can create your own packages and contribute them back to the ecosystem

R



Front Ends *(Text Editors, IDEs, GUIs)*

Vim Emacs + ESS Eclipse + StatET Rstudio Deducer

Understanding How is Data Organized: Key Terms and Technologies

- **Database**: A single table or a collection of related tables
- **Database management systems (DBMS)**: Sometimes called “database software”; software for creating, maintaining, and manipulating data
- **Structured query language (SQL):** A language used to create and manipulate databases
- **Database administrator (DBA):** Job title focused on directing, performing, or overseeing activities associated with a database or set of databases
  - Includes database design, creation, implementation, maintenance, backup and recovery, policy setting and enforcement, and security

Understanding How is Data Organized: Key Terms and Technologies

- **Table or file**: A list of data, arranged in columns (fields) and rows (records)
- **Column or field**: A column in a database table. Columns represent each category of data contained in a record (e.g., first name, last name, ID number, data of birth)

Understanding How is Data Organized: Key Terms and Technologies

- **Row or record or tuple**: A row in a database table. Records represent a single instance of whatever the table keeps track of (e.g., student, faculty, course title)
- **Key**: A field or combination of fields used to uniquely identify a record, and to relate separate tables in a database. Examples include social security number, customer account number, or student ID
- **Relational database**: The most common standard for expressing databases, whereby tables (files) are related based on common keys

Where Does Data Come From?

- For organizations that sell directly to their customers, transaction processing systems represent a fountain of potentially insightful data
  - **Transaction processing systems (TPS)**: A system that records a transaction (some form of business-related exchange), such as a cash register sale, ATM withdrawal, or product return
  - **Transaction**: Some kind of business exchange
  - The cash register is the primary source that feeds data to the TPS
  - TPS can generate a lot of bits, it’s sometimes tough to match this data with a specific customer

Where Does Data Come From?

- Enterprise software (CRM, SCM, and ERP)
  - Firms set up systems to gather additional data beyond conventional purchase transactions or Web site monitoring
  - CRM or customer relationship management systems are used to empower employees to track and record data at nearly every point of customer contact
  - Supply chain management (SCM) and enterprise resource planning (ERP) systems touch every aspect of the value chain

Where Does Data Come From?

- Surveys
  - Firms supplement operational data with additional input from surveys and focus groups
  - Direct surveys can tell you what your cash register can’t
  - Many CRM products have survey capabilities that allow for additional data gathering at all points of customer contact

Where Does Data Come From?

- External sources
  - If your firm has partners that sell products for you, then you'll likely rely heavily on data collected by others
  - Data bought from sources available to all might not yield competitive advantage on its own. But it can provide key operational insight for increased efficiency and cost savings

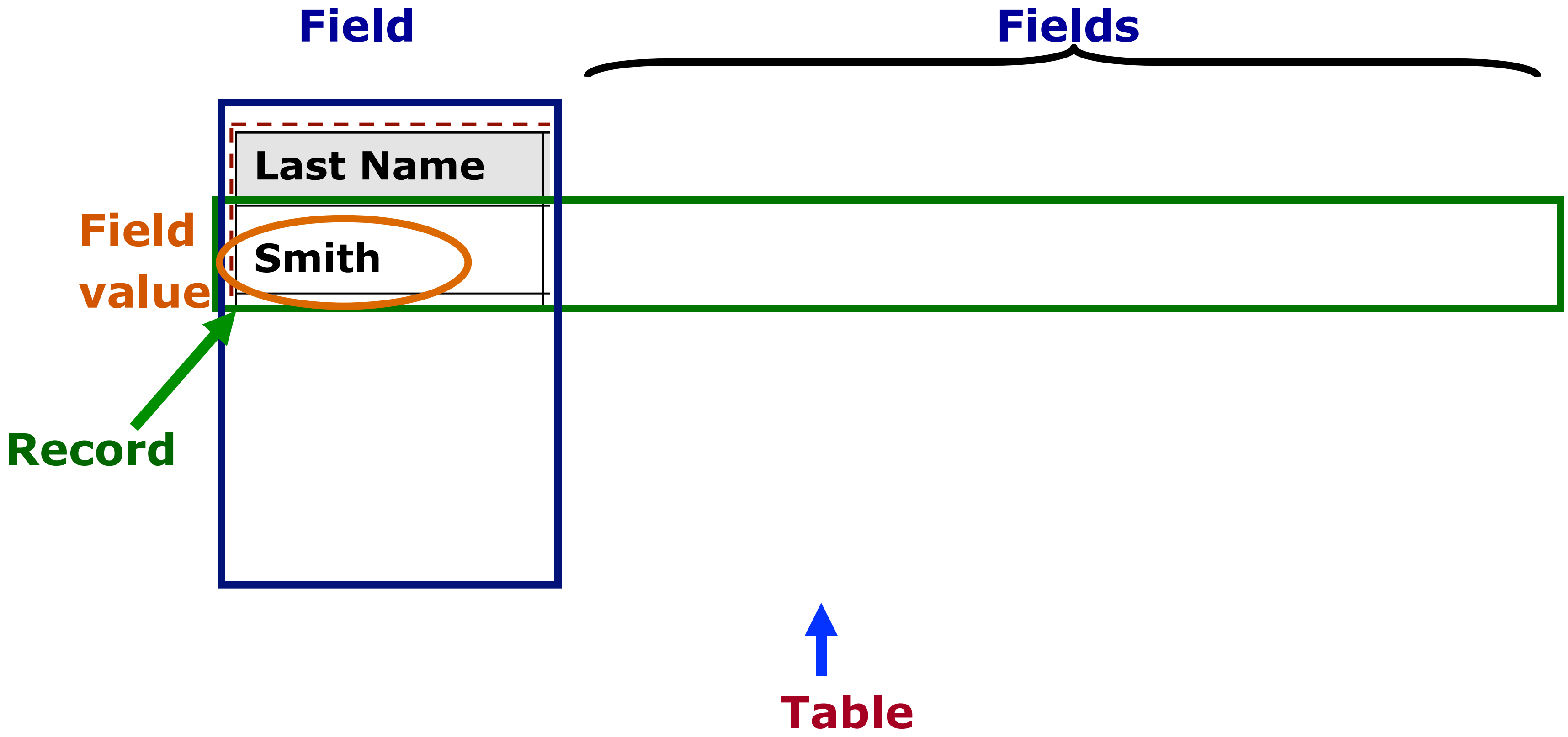
Data Rich, Information Poor

- Many organizations are data rich but information poor
- Factors holding back information advantage
  - **Legacy system**: Older information systems that are often incompatible with other systems, technologies, and ways of conducting business
  - Most transactional databases aren't set up to be simultaneously accessed for reporting and analysis

Understanding How Data is Organized

What is a table and what does it look like?

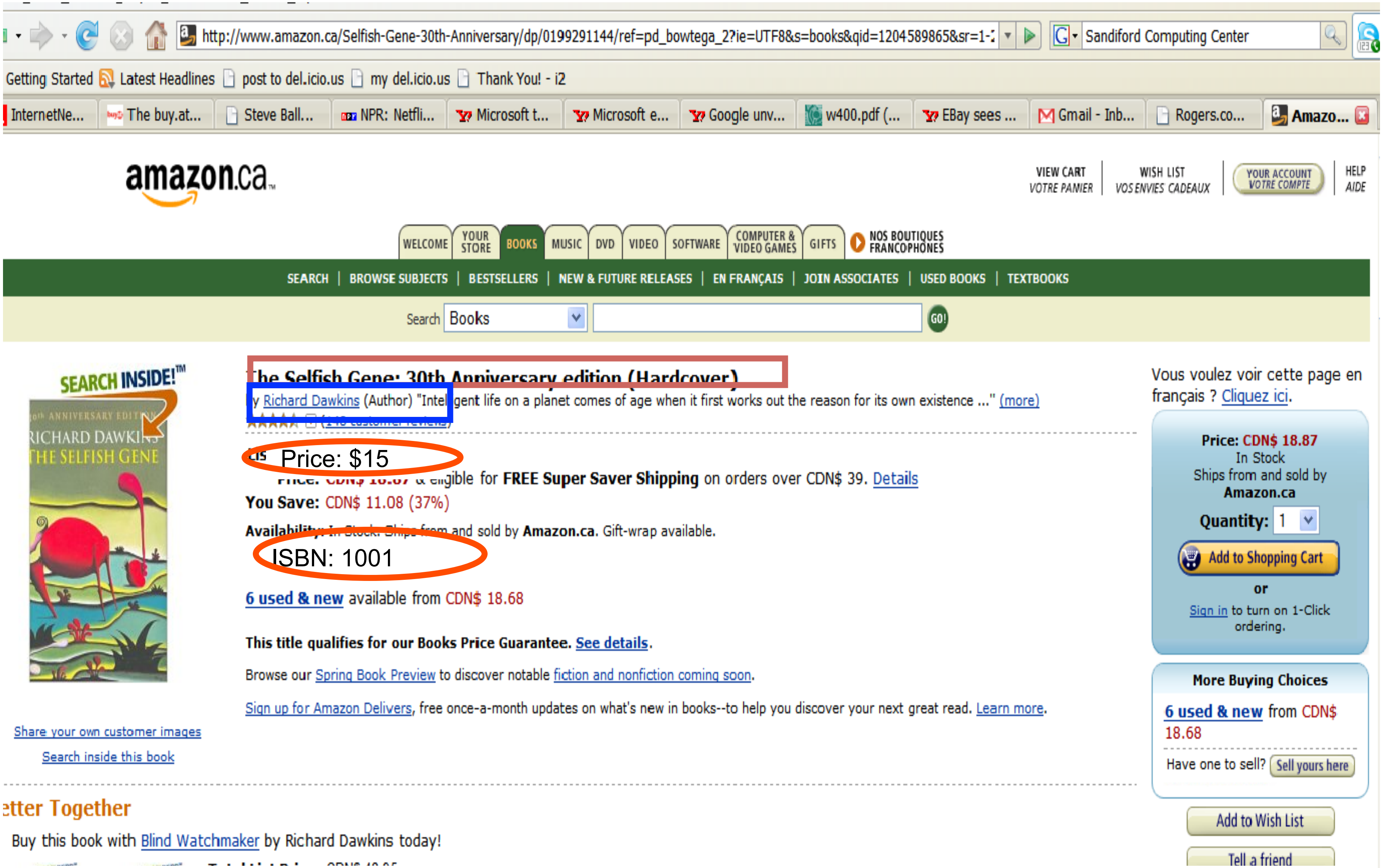
Composed of records which are composed of fields/attributes.



Example: Database for Amazon

- What are important “things” that Amazon needs to keep track of?

Let’s focus on tracking information about books



The Selfish Gene: 30th Anniversary edition (Hardcover)

by Richard Dawkins (Author) "Intelligent life on a planet comes of age when it first works out the reason for its own existence ..." (more)

(140 customer reviews)

Price: $15

Price: CDN$ 18.87 & eligible for FREE Super Saver Shipping on orders over CDN$ 39. Details

You Save: CDN$ 11.08 (37%)

Availability: In Stock. Ships from and sold by Amazon.ca. Gift-wrap available.

ISBN: 1001

6 used & new available from CDN$ 18.68

This title qualifies for our Books Price Guarantee. See details.

Browse our Spring Book Preview to discover notable fiction and nonfiction coming soon.

Sign up for Amazon Delivers, free once-a-month updates on what's new in books--to help you discover your next great read. Learn more.

Share your own customer images

Search inside this book

Vous voulez voir cette page en français ? Cliquez ici.

Price: CDN$ 18.87
In Stock
Ships from and sold by
Amazon.ca

Quantity: 1

or

Sign in to turn on 1-Click ordering.

More Buying Choices

6 used & new from CDN$ 18.68

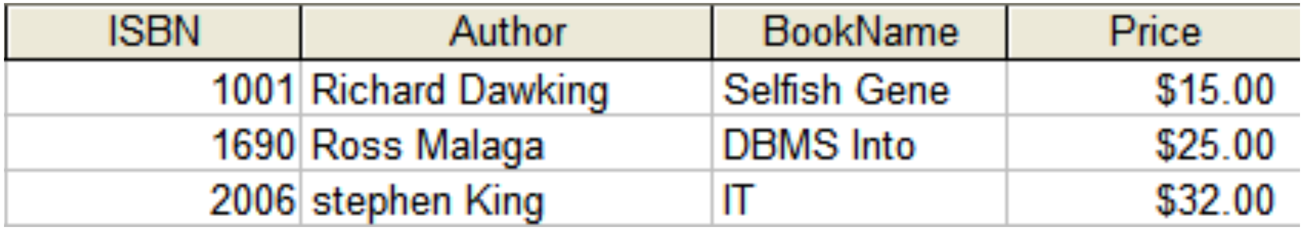
Have one to sell?

etter Together

Buy this book with Blind Watchmaker by Richard Dawkins today!

Note: ISBN number shown here for this book is not real (it is a made up number to make example simple)

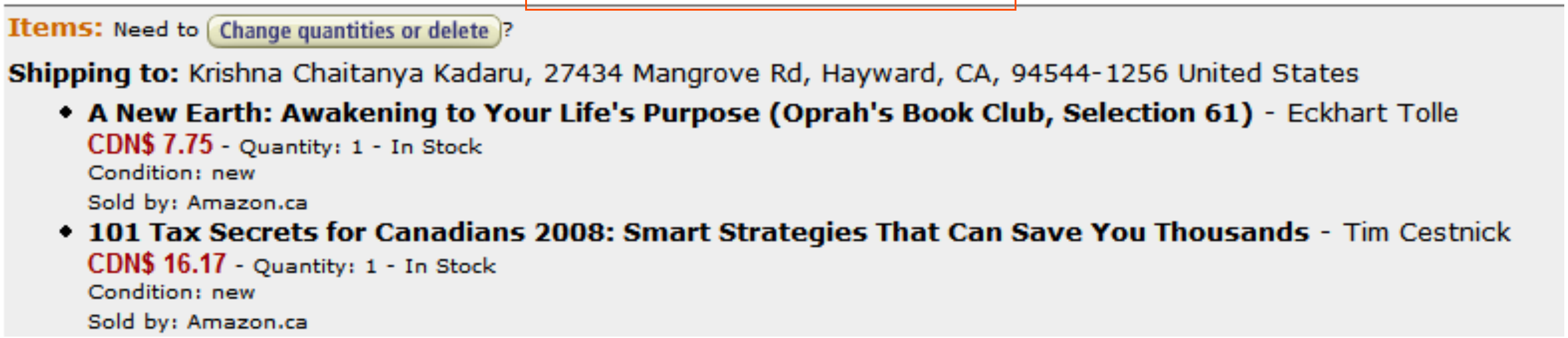
How would you store book related data if you were using Excel?



ISBN	Author	BookName	Price
1001	Richard Dawking	Selfish Gene	$15.00
1690	Ross Malaga	DBMS Into	$25.00
2006	stephen King	IT	$32.00

What if we also want to track “orders” placed by customers?

A sample order



Items: Need to Change quantities or delete?

Shipping to: Krishna Chaitanya Kadaru, 27434 Mangrove Rd, Hayward, CA, 94544-1256 United States

- **A New Earth: Awakening to Your Life's Purpose (Oprah's Book Club, Selection 61)** - Eckhart Tolle
  CDN$ 7.75 - Quantity: 1 - In Stock
  Condition: new
  Sold by: Amazon.ca
- **101 Tax Secrets for Canadians 2008: Smart Strategies That Can Save You Thousands** - Tim Cestnick
  CDN$ 16.17 - Quantity: 1 - In Stock
  Condition: new
  Sold by: Amazon.ca

Very Simplified Order Form

Order # ___
Date: _____

Send to: Customer’s name
Phone

Detail:

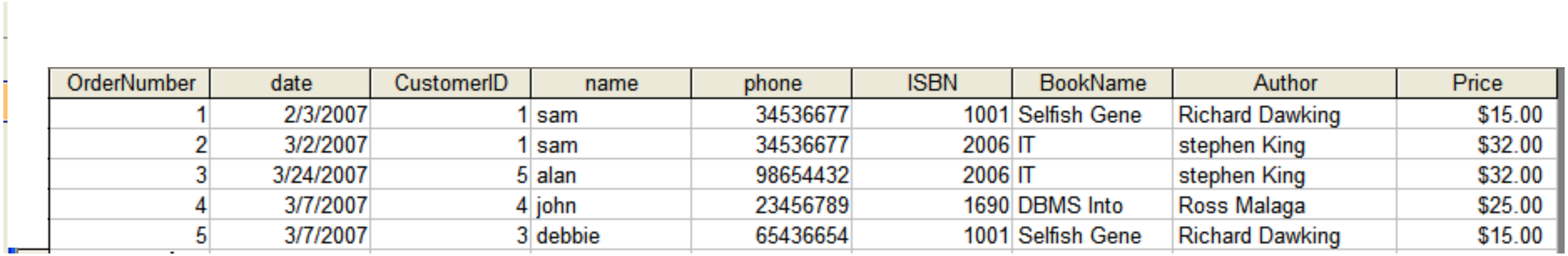
ISBN	*Book Name*	*Author*	*Price*
__________	____________	____________	_________

Recap: Information that we want to track

- **Order related**
  - Order Number, Date
- **Customer related**
  - Customer ID, Name, Phone
- **Books related**
  - ISBN, Name, Author, Price

How would you store such data if you were using Excel?

Organizing Order Information in Excel



OrderNumber	date	CustomerID	name	phone	ISBN	BookName	Author	Price
1	2/3/2007	1	sam	34536677	1001	Selfish Gene	Richard Dawking	$15.00
2	3/2/2007	1	sam	34536677	2006	IT	stephen King	$32.00
3	3/24/2007	5	alan	98654432	2006	IT	stephen King	$32.00
4	3/7/2007	4	john	23456789	1690	DBMS Into	Ross Malaga	$25.00
5	3/7/2007	3	debbie	65436654	1001	Selfish Gene	Richard Dawking	$15.00

Problems?

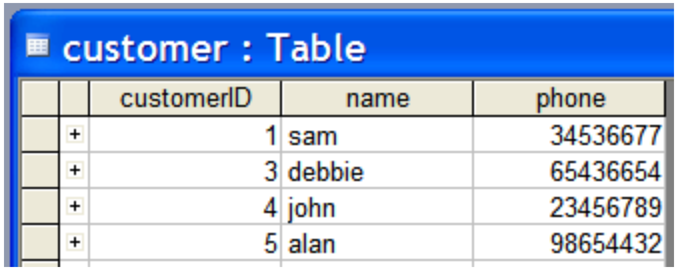
1. **Adding** a potential customer who has not ordered yet
   Adding a book recently received from supplier (i.e., it has not been ordered by any customer)
2. **Deleting** an order (such as order # 4 or order # 5)
3. **Modifying** an attribute (Changing price of book named “selfish gene”)

Cause?

Data is not structured properly (i.e., un-normalized)

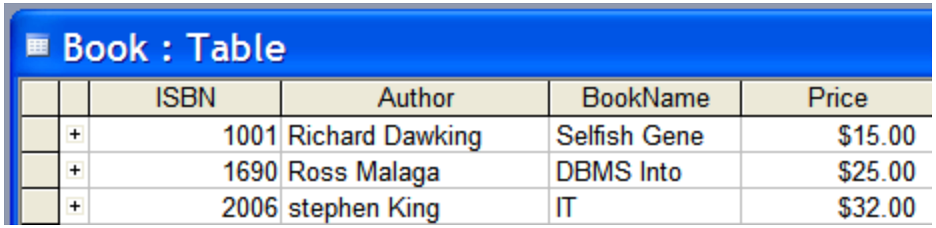
Relational Database Approach

Create a series of **logically related** **two-dimensional tables** to store their information



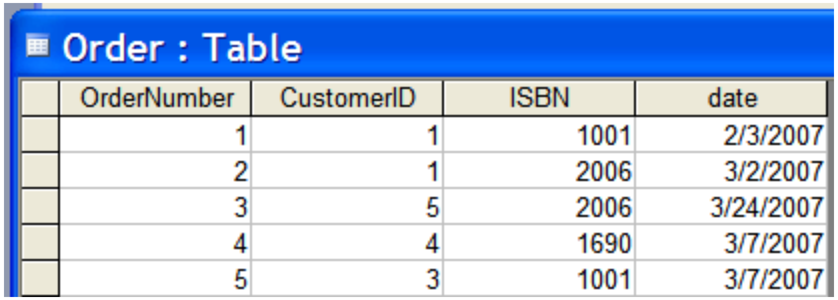
customer : Table

	customerID	name	phone
+	1	sam	34536677
+	3	debbie	65436654
+	4	john	23456789
+	5	alan	98654432



Book : Table

	ISBN	Author	BookName	Price
+	1001	Richard Dawking	Selfish Gene	$15.00
+	1690	Ross Malaga	DBMS Into	$25.00
+	2006	stephen King	IT	$32.00



Order : Table

OrderNumber	CustomerID	ISBN	date
1	1	1001	2/3/2007
2	1	2006	3/2/2007
3	5	2006	3/24/2007
4	4	1690	3/7/2007
5	3	1001	3/7/2007

We have **3 tables** – a table for book related data, another for customer related data, and finally a table for order related data

How are these “logically” related?

Connecting tables together

- **1. Each Table should have a Primary Key**
  - **Primary keys**
    - A field/attribute (or group of fields/attributes in some cases) that **uniquely** identify each record/entity in a table
    - Examples: Customer ID, ISBN, Order#

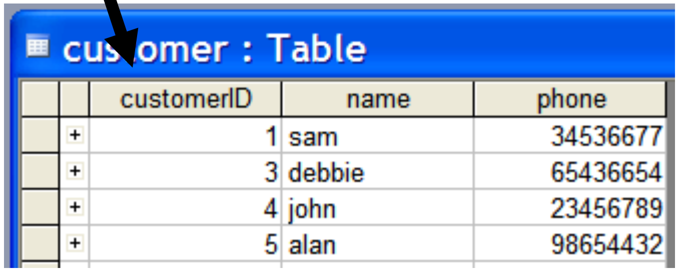
- **2. Tables are connected using Foreign Keys**
  - **Foreign keys**
    - A field that is a primary key in one table and appears in a different table (may appear as a part of the primary key)
    - Examples: Customer ID in **Orders** table
    - Another example: ???

Note: Primary Key (PK) is identified by underlining appropriate field/s.

Logical Structure of the database:

Each Table should have a **Primary Key**

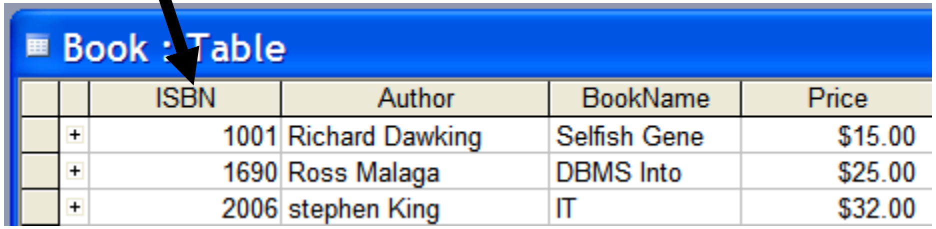
Primary Key



customer : Table

customerID	name	phone
1	sam	34536677
3	debbie	65436654
4	john	23456789
5	alan	98654432

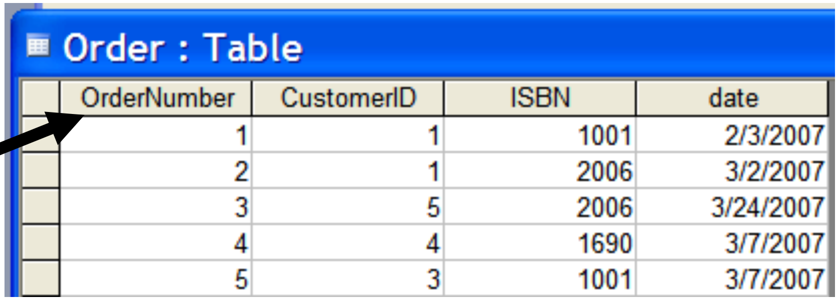
Primary Key



Book : Table

ISBN	Author	BookName	Price
1001	Richard Dawking	Selfish Gene	$15.00
1690	Ross Malaga	DBMS Into	$25.00
2006	stephen King	IT	$32.00

Primary Key

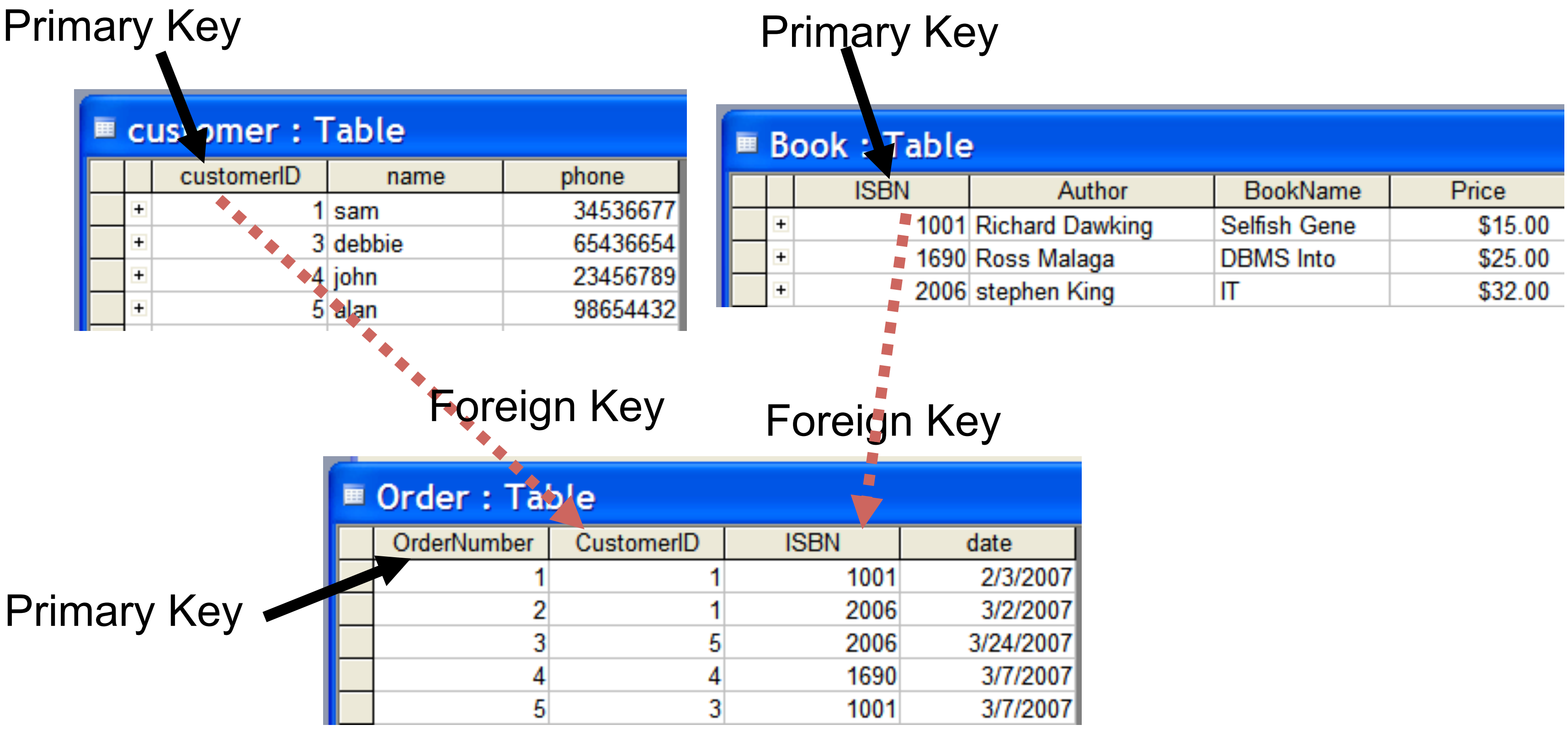


Order : Table

OrderNumber	CustomerID	ISBN	date
1	1	1001	2/3/2007
2	1	2006	3/2/2007
3	5	2006	3/24/2007
4	4	1690	3/7/2007
5	3	1001	3/7/2007

Logical Structure of the database:

Tables are connected using **Foreign Keys**



Primary Key

customer : Table

	customerID	name	phone
+	1	sam	34536677
+	3	debbie	65436654
+	4	john	23456789
+	5	alan	98654432

Primary Key

Book : Table

	ISBN	Author	BookName	Price
+	1001	Richard Dawking	Selfish Gene	$15.00
+	1690	Ross Malaga	DBMS Into	$25.00
+	2006	stephen King	IT	$32.00

Foreign Key

Foreign Key

Primary Key

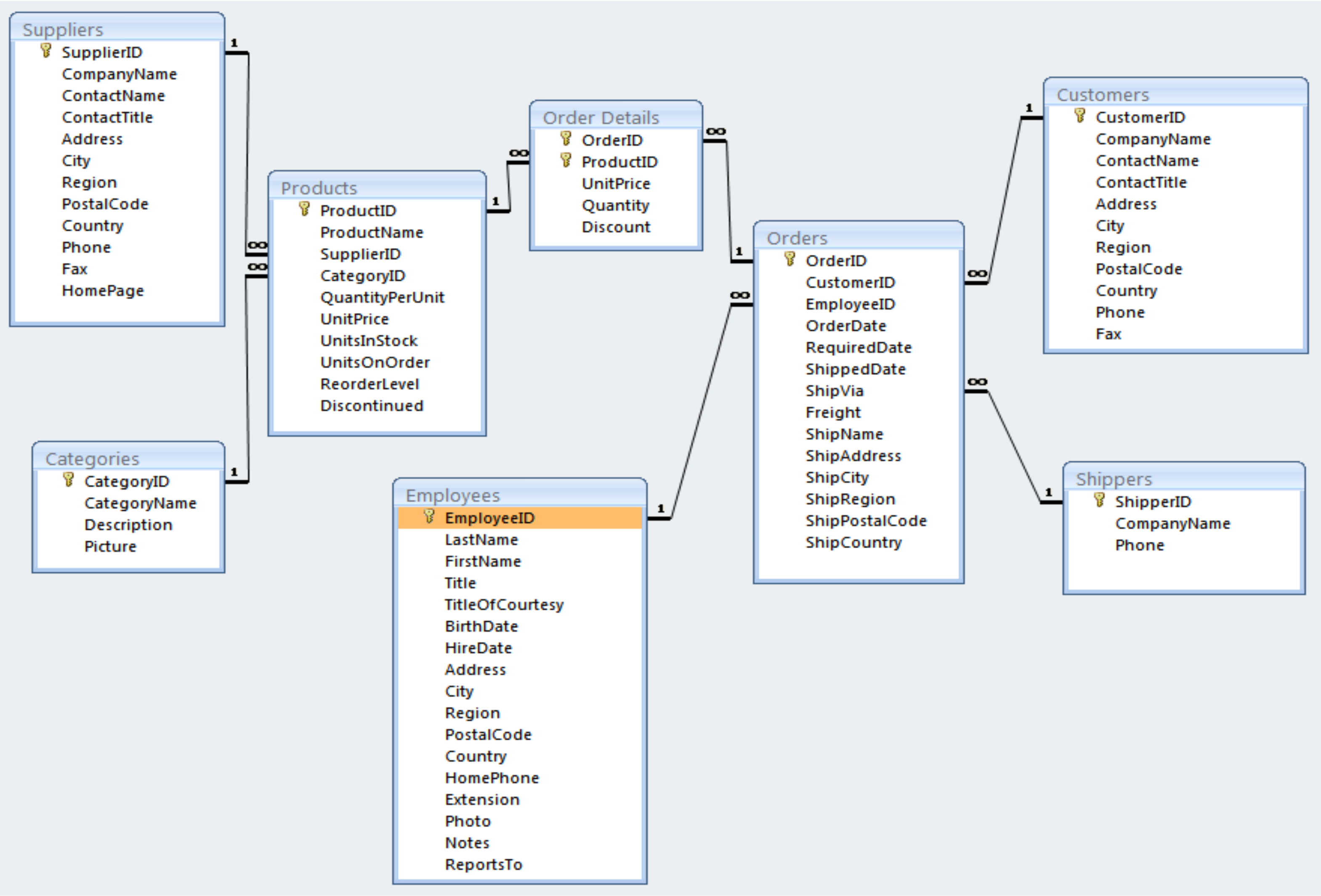
Order : Table

OrderNumber	CustomerID	ISBN	date
1	1	1001	2/3/2007
2	1	2006	3/2/2007
3	5	2006	3/24/2007
4	4	1690	3/7/2007
5	3	1001	3/7/2007

Text Representation of Tables →

Customer (CustomerID, Name, Phone)
Book (ISBN, Author, BookName, Price)
Order (OrderNumber, CustomerID, ISBN, Date)

Tables are connected by creating relationships



Suppliers
- SupplierID
- CompanyName
- ContactName
- ContactTitle
- Address
- City
- Region
- PostalCode
- Country
- Phone
- Fax
- HomePage

Products
- ProductID
- ProductName
- SupplierID
- CategoryID
- QuantityPerUnit
- UnitPrice
- UnitsInStock
- UnitsOnOrder
- ReorderLevel
- Discontinued

Order Details
- OrderID
- ProductID
- UnitPrice
- Quantity
- Discount

Customers
- CustomerID
- CompanyName
- ContactName
- ContactTitle
- Address
- City
- Region
- PostalCode
- Country
- Phone
- Fax

Orders
- OrderID
- CustomerID
- EmployeeID
- OrderDate
- RequiredDate
- ShippedDate
- ShipVia
- Freight
- ShipName
- ShipAddress
- ShipCity
- ShipRegion
- ShipPostalCode
- ShipCountry

Categories
- CategoryID
- CategoryName
- Description
- Picture

Employees
- EmployeeID
- LastName
- FirstName
- Title
- TitleOfCourtesy
- BirthDate
- HireDate
- Address
- City
- Region
- PostalCode
- Country
- HomePhone
- Extension
- Photo
- Notes
- ReportsTo

Shippers
- ShipperID
- CompanyName
- Phone

Each primary key - foreign key pair represents a relationship.

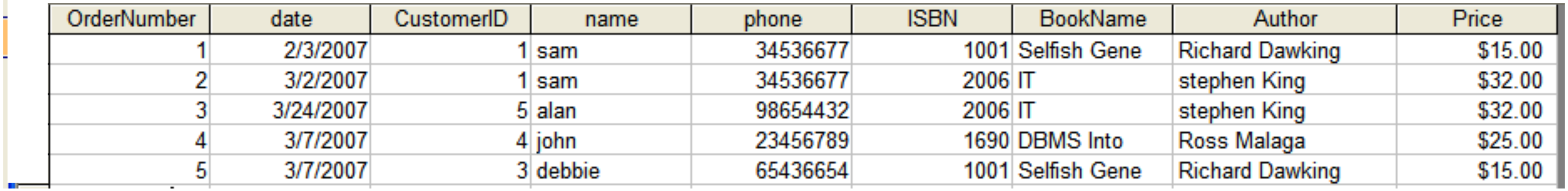
Normalization

- Problems/Anomalies arise if data is not structured properly (i.e., un-normalized)
- **Normalization** is a method for analyzing and reducing a relational database to its most streamlined form for:
  - Minimum redundancy
  - Maximum data integrity
  - Best processing performance
- Normalized data is when attributes in the table depend only on the primary key.

- **How to create normalized tables is beyond the scope of this course (covered in INSY 333 and INSY 437).**

Normalization

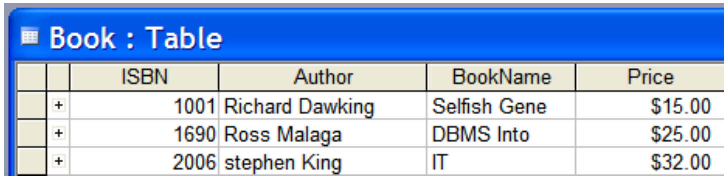
Un-Normalized Table



OrderNumber	date	CustomerID	name	phone	ISBN	BookName	Author	Price
1	2/3/2007	1	sam	34536677	1001	Selfish Gene	Richard Dawking	$15.00
2	3/2/2007	1	sam	34536677	2006	IT	stephen King	$32.00
3	3/24/2007	5	alan	98654432	2006	IT	stephen King	$32.00
4	3/7/2007	4	john	23456789	1690	DBMS Into	Ross Malaga	$25.00
5	3/7/2007	3	debbie	65436654	1001	Selfish Gene	Richard Dawking	$15.00

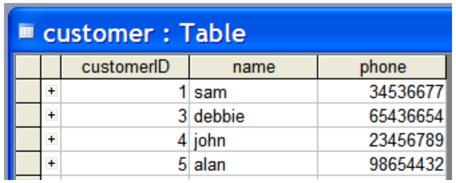
Normalized Tables

Customer (CustomerID, name, phone)
Book (ISBN, Author, BookName, Price)
Order (OrderNumber, date, CustomerID, ISBN)



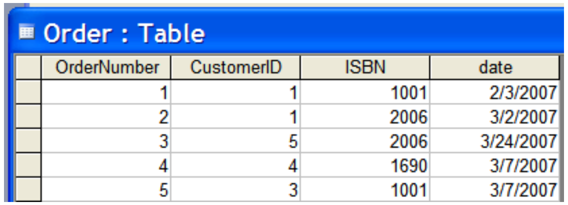
Book : Table

	ISBN	Author	BookName	Price
+	1001	Richard Dawking	Selfish Gene	$15.00
+	1690	Ross Malaga	DBMS Into	$25.00
+	2006	stephen King	IT	$32.00



customer : Table

	customerID	name	phone
+	1	sam	34536677
+	3	debbie	65436654
+	4	john	23456789
+	5	alan	98654432



Order : Table

OrderNumber	CustomerID	ISBN	date
1	1	1001	2/3/2007
2	1	2006	3/2/2007
3	5	2006	3/24/2007
4	4	1690	3/7/2007
5	3	1001	3/7/2007

Data Scientists and Relational Databases

Data scientists need to be able to:

- Look at a relational database and understand how data is organized
- Select and extract data from multiple tables using SQL
- Perform in database calculations using SQL

Abstraction: SQL Enables Relational Algebra Calculations

Operations always create new relations

Operations (filter, refine)

- Selection
- Projection
- Cartesian product (join)
- Set union
- Set Difference
- Rename

Different Select Statements

Basic Selection

```
SELECT * FROM batting;
```

Basic Selection of Big Table

```
SELECT * FROM batting limit 50;
```

Different Select Statements

Basic Selection of Different Columns

SELECT H, AB, 2B, 3B FROM batting;

Basic Selection of Specific Rows (here from a specific year)

SELECT * FROM batting where yearID=1950;

Basic Selection of Specific Rows (here from a specific year)

SELECT * FROM batting where yearID=1950 and teamID = “KCA”;

Different Select Statements

Calculate a new field

```
SELECT *, H/AB AS AVG
        , (H+BB+HBP)/(AB+BB+HBP+SF) AS OBP
FROM batting;
```

Different Select Statements

Calculate a new field

```
SELECT playerid year from batting where HR > 60;
SELECT count(playerID) from batting where HR
>60;
SELECT count(DISTINCT playerID) from batting
where HR >60;
```

Different Select Statements

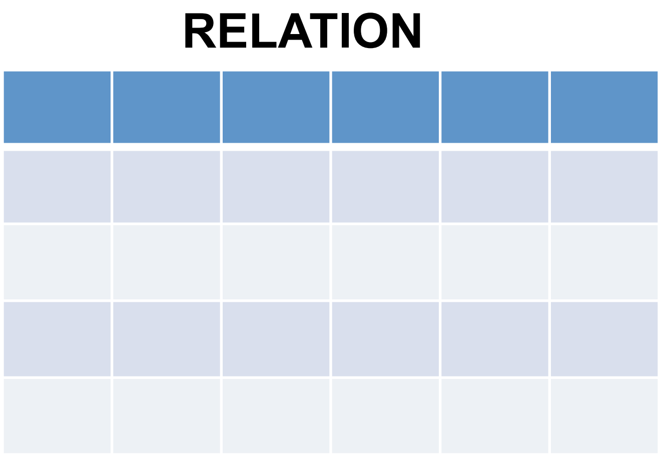
Aggregation

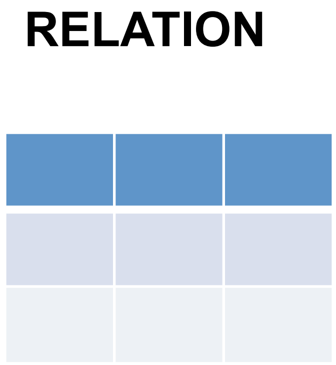
```
select teamid, yearID, SUM(salary) from Salaries group by teamid, yearid;
select yearID, SUM(salary) from Salaries group by yearid;
select yearID, AVG(salary) from Salaries group by yearid;
```

Note how you have to perform aggregations on some variables

Select with SQL

- The product of any selection is another relation





- This means that relations can be nested (subselect).

Different Select Statements

Subselect

```
SELECT playerID, yearID, teamID, HR FROM
batting where exists (SELECT playerID,IPOuts/3
as IP from pitching where IPOuts >300) order by
HR desc;
```

Here we are looking for pitchers who also can hit!

Different Select Statements

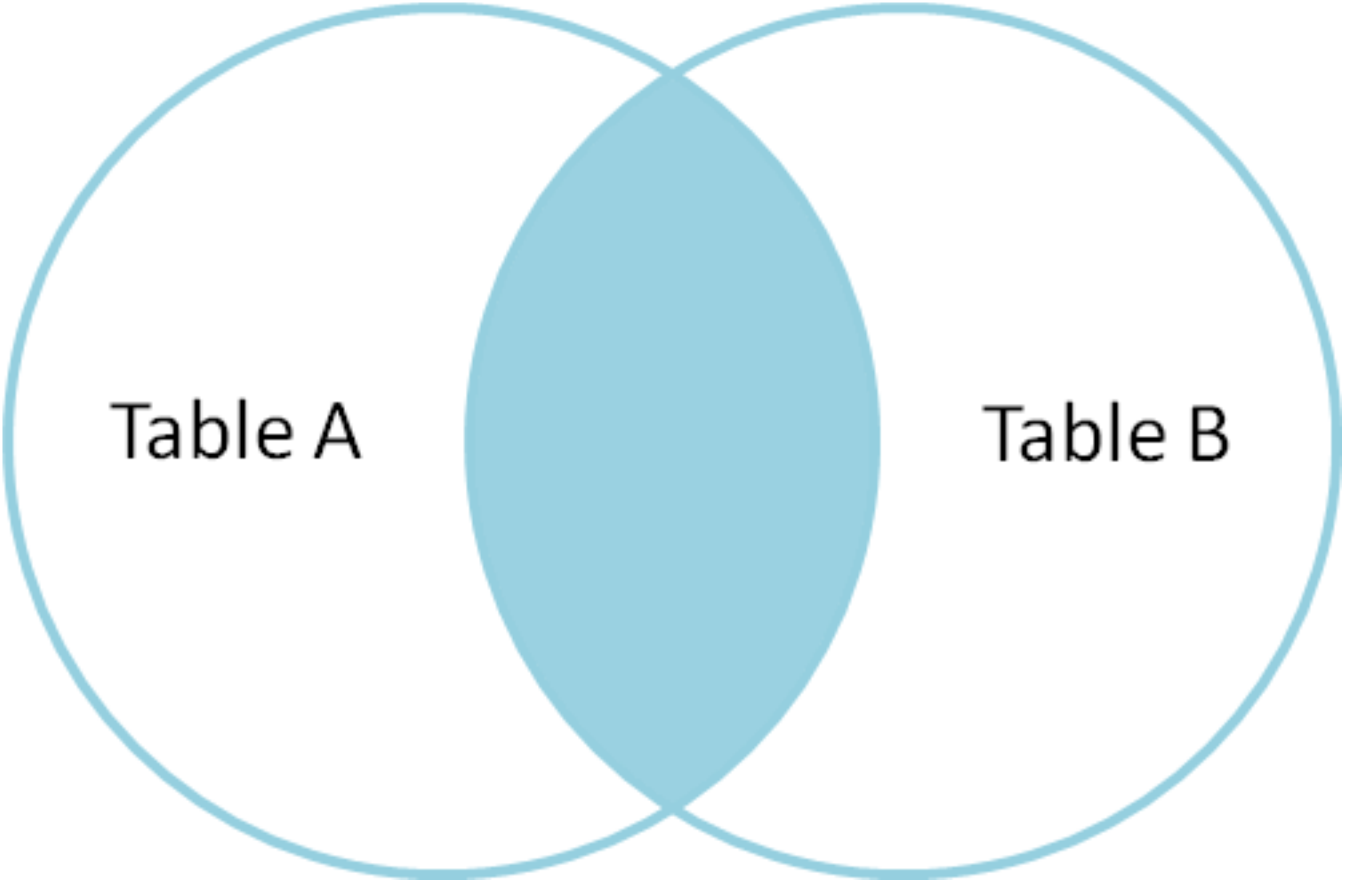
Joins

```
SELECT p.playerID, m.nameFirst, m.nameLast,
p.W p.L FROM pitching p, master m WHERE
p.playerID = m.playerID;
```

Here we are including the first and last name

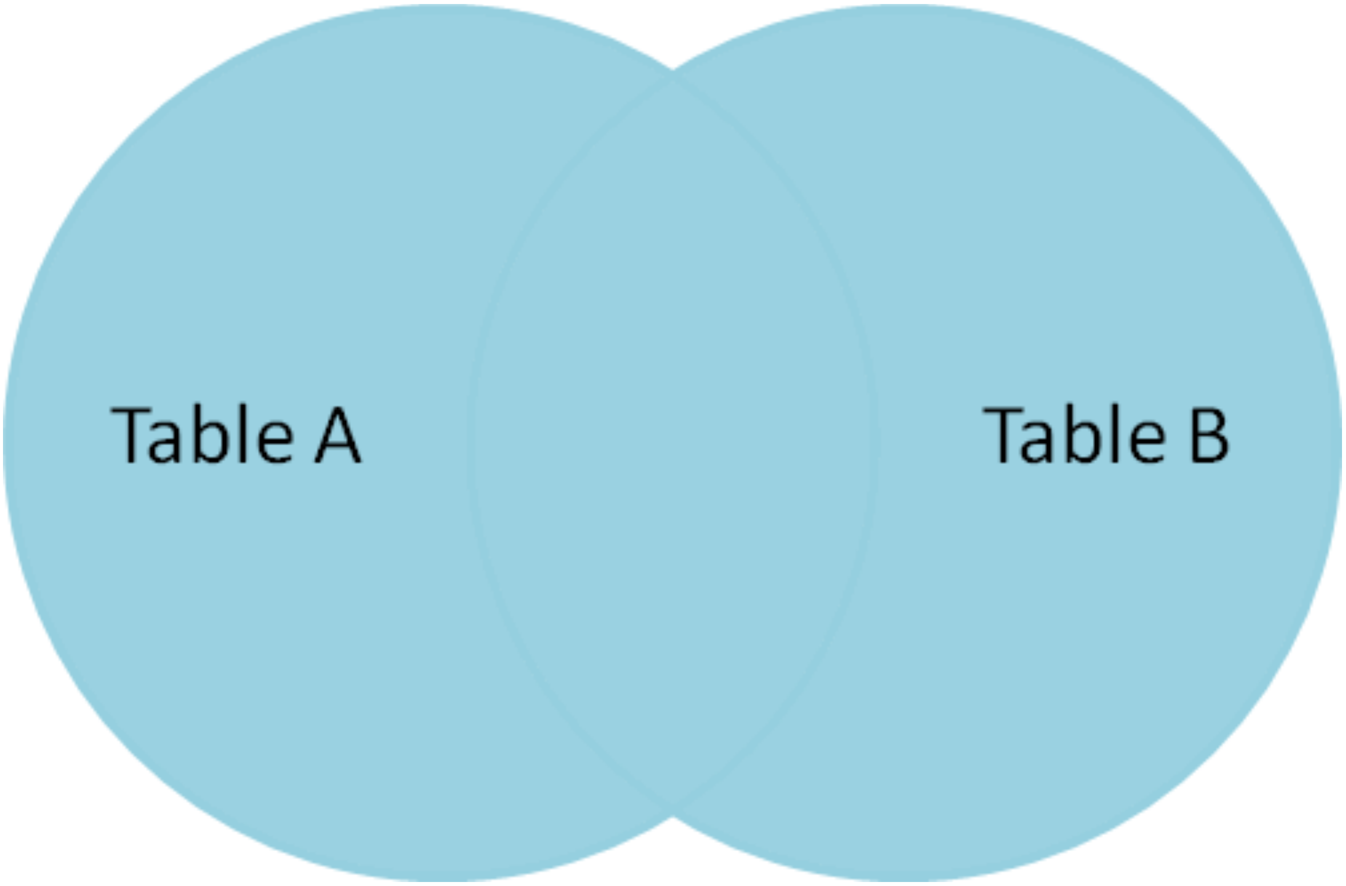
Joins, Visually

INNER JOIN



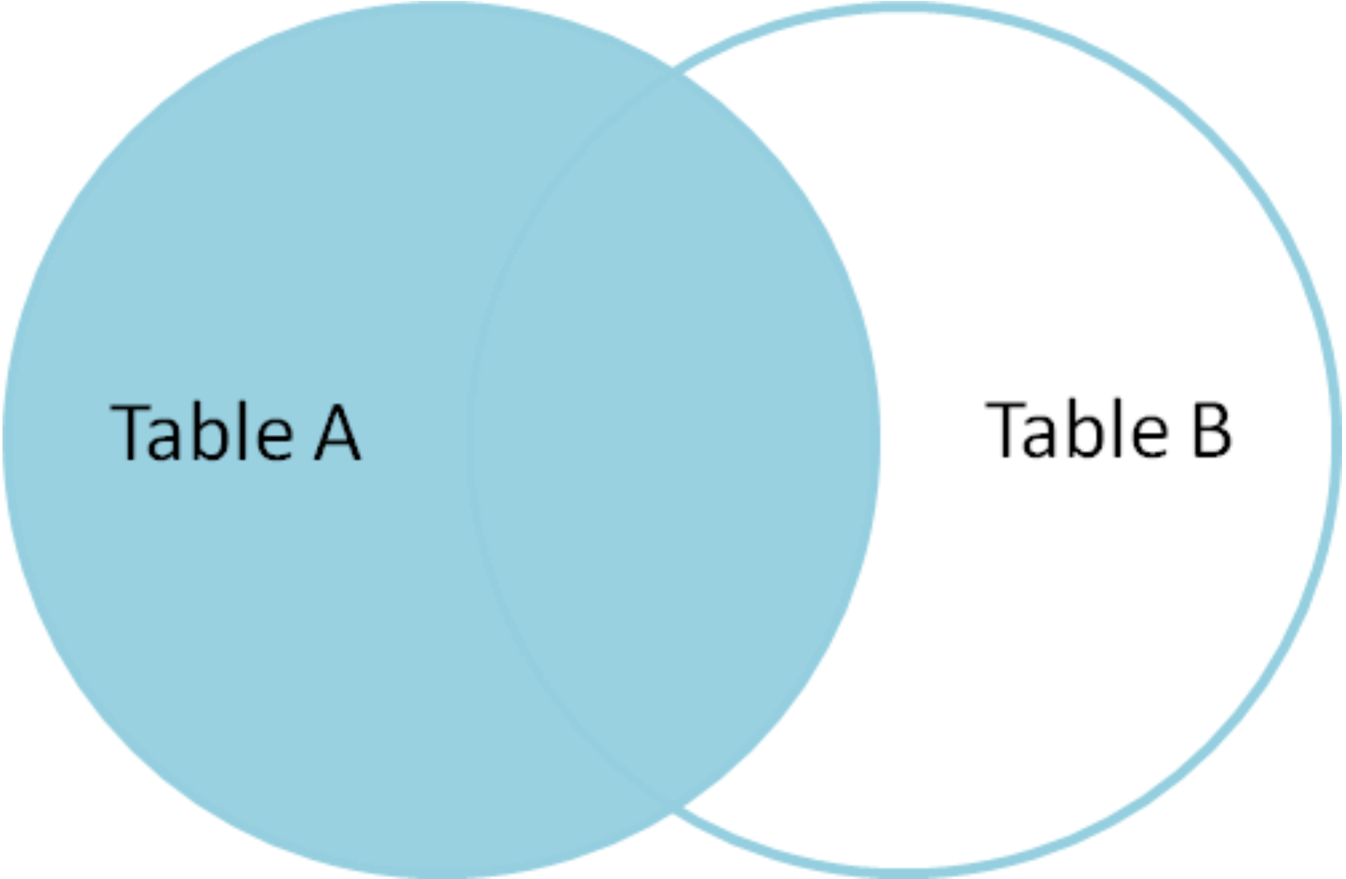
Joins, Visually

Full Outer Join



Joins, Visually

Left Outer Join



More advanced SQL Continued in SQL Lab

Level of Analysis and Aggregation

- If you were going to try to analyze those factors that drive team success, why couldn't you include information directly from the batting table?

Level of Analysis and Aggregation

- Easier to include higher level factors in lower levels of analysis than the opposite
  - Leagues
  - Division
  - Team
  - Players
- To include player level variables in team level analyses you have to aggregate them

Level of Analysis and Aggregation

- Salary Analysis
  - In a player analysis of it would be relevant and appropriate to include “dummy variables” indicating the teams they are playing for
  - Yankees are likely to earn more than Pirates