IST 3420: Introduction to Data Science and Management

Langtao Chen, Fall 2017

3. Data Basics

Reading Assignment 6 (due Sep 3)

- Stevens, S. S. (1946). On the Theory of Scales of Measurement. *Science*, *103*(2684), 677-680.

- Read "HTML Tutorial" from http://www.w3schools.com/html/default.asp
  - You need to read from "HTML HOME" to "HTML Tables".

Agenda

- Data, Dataset, and Scales of Measurement
- Data Collection
- Working with CSV
- Working with Rational Database
- Working with HTML
- Working with XML and JSON
- Working with APIs

Data and Data Set

- Data are the facts collected, analyzed, and interpreted.
- The data collected in a particular data science project are commonly referred to as a data set.

Data Set: Elements, Variables, and Observations

- Elements/subjects: entities of interest
- Variables: characteristic of elements
- Observation: the set of measurements obtained for an element

Variables

car	mpg	cyl	hp	wt
Mazda RX4	21	6	110	2.62
Mazda RX4 Wag	21	6	110	2.875
Datsun 710	22.8	4	93	2.32
Hornet 4 Drive	21.4	6	110	3.215
Hornet Sportabout	18.7	8	175	3.44
Valiant	18.1	6	105	3.46

Element Names

Observations

Scales of Measurement

- Scale/level of measurement determines:
  - the amount of information contained in data
  - data summarization and analysis methods that are appropriate

- Four types of scales
  - Nominal
  - Ordinal

    Qualitative / categorical data
  - Interval
  - Ratio

    Quantitative data

Nominal Scale

- Numerical values are just names or labels of the attribute
  - Ordering of these values is meaningless
  - No mathematical calculation (+, -, *, /) applicable

- For example:
  - Gender (1 = “Male”, 0 = “Female”)
  - Student ID (1,2,3...)
  - Department (1 = “BIT”, 2 = “CS”...)
  - Zip code (65401, 65402...)

Ordinal Scale

- Attributes can be ranked/ordered.

- For example:
  - Football team rank (1st, 2nd, 3rd...)
  - Customer rating (1 = "Bad", 2 = "OK", 3 = "Excellent")

Interval Scale

- Have all characteristics of ordinal scale
- Distance between attributes does have meaning.
- Ratios are not meaningful.

- For example:
  - Temperature
    - The distance from 40 – 60 is same as the distance from 60 – 80
    - 80 cannot be said as twice hot as 40
  - SAT Score
  - GMAT Score

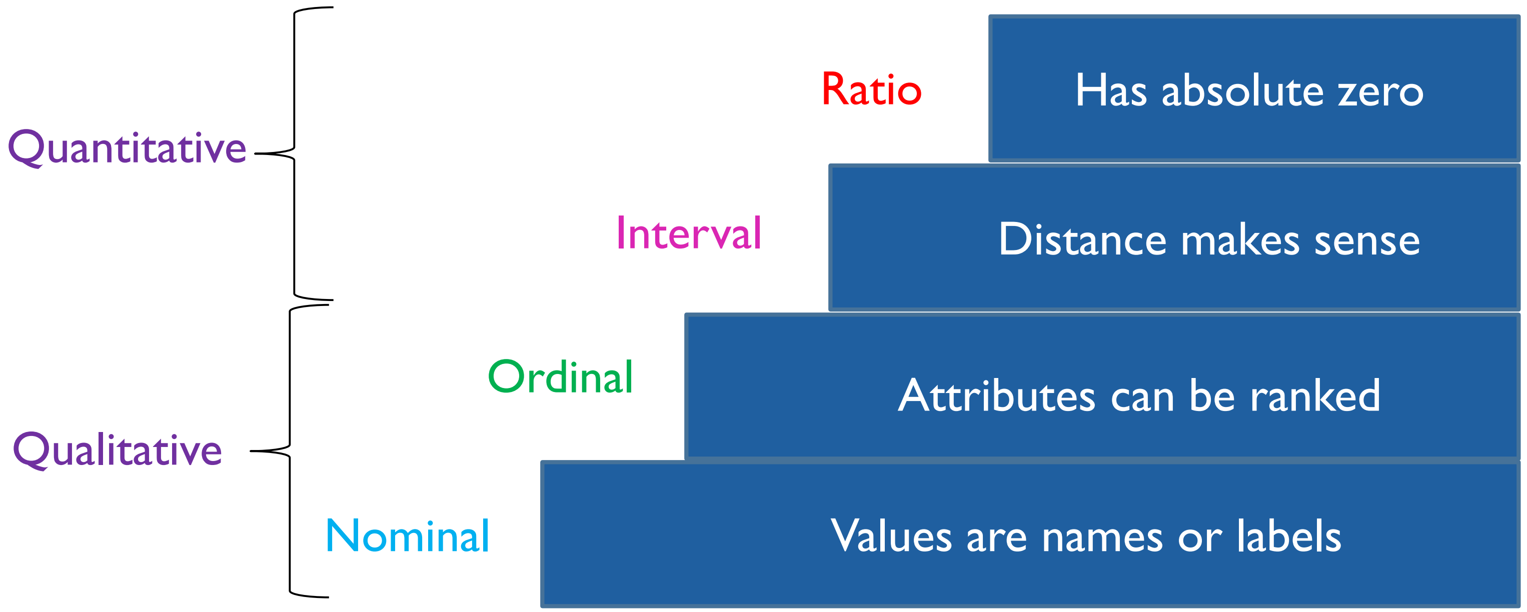
Ratio Scale

- Have all characteristics of interval scale
- A ratio of two values is meaningful.
- An absolute zero is meaningful.

- For example:
  - Weight
  - Height
  - Distance
  - Number of visits
  - Credit hours earned

Hierarchy of Measurement Scales

- A higher level scale contains all properties of its lower scale.
- From lower to higher levels, analysis tends to be more comprehensive. Improper use of lower level scales suffers information loss in the data
- In general, we prefer a higher scale of measurement than a lower one.

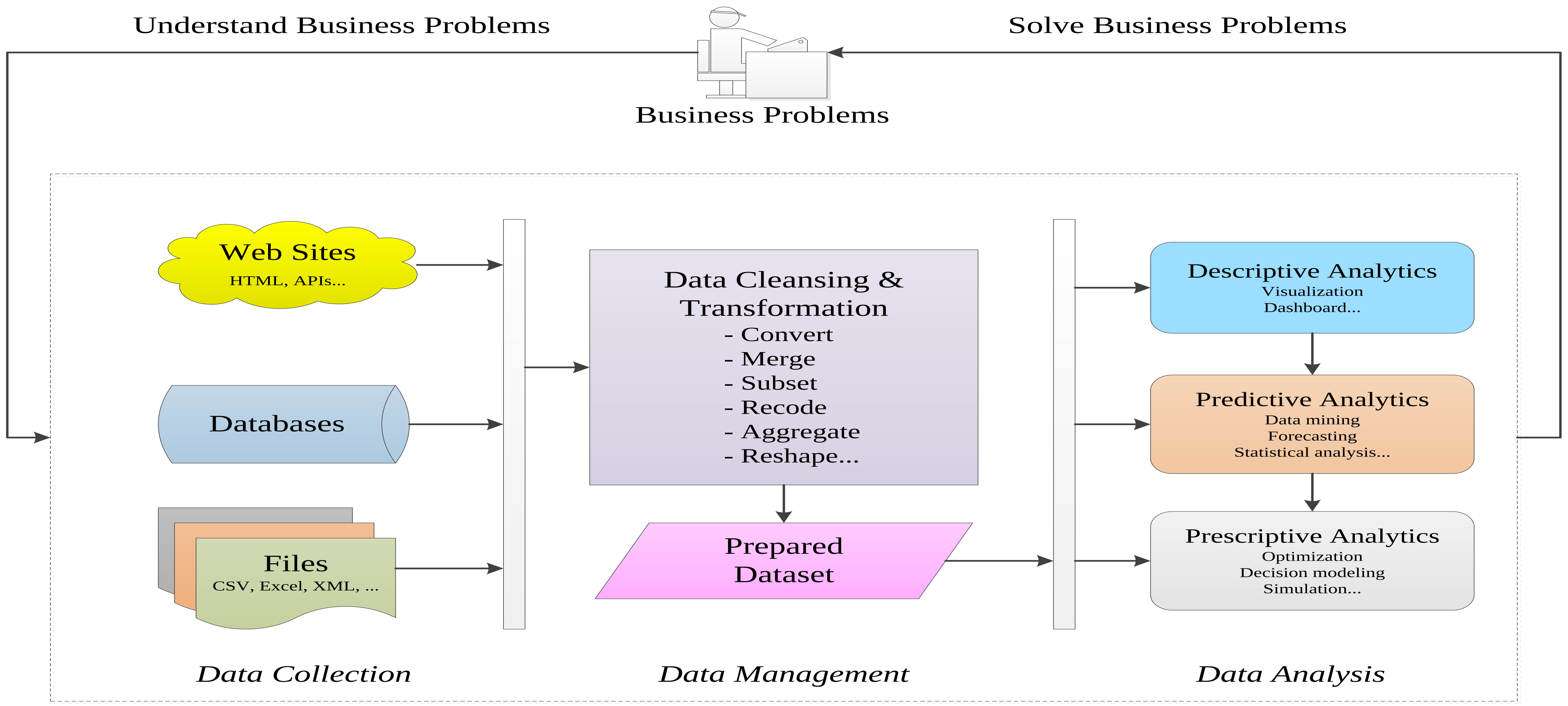


Agenda

- Data, Dataset, and Scales of Measurement
- Data Collection
- Working with CSV
- Working with Rational Database
- Working with HTML
- Working with XML and JSON
- Working with APIs

Data Collection for Data Science

- Data collection is the first step in data science procedures.



Data Sources

- Primary data: collect first-hand data through experimental or observational studies (e.g., survey)

- Secondary data: reuse existing data
  - Database or data warehouse
  - Dataset collected by someone else
  - Data downloaded from organizations such as government agencies and industry associations
  - Digital trace data recorded by computer systems
  - Web server logs
  - Internet social media and user-generated content
  - …

- In this course, we focus on the management and analysis of existing data.

Major Contents to Cover

- General principles for data management
- Common data representation structures such as database, CSV, HTML, XML, JSON
- Commonly used data collection methods

Common Data Collection Methods

- CSV
- Relational Database
- XML and JSON
- Web Scraping (copy and paste is inefficient)
  - HTML
  - API

Web scraping (web harvesting or web data extraction) is a computer software technique of extracting information from websites.
---- Wikipedia (https://en.wikipedia.org/wiki/Web_scraping)

Agenda

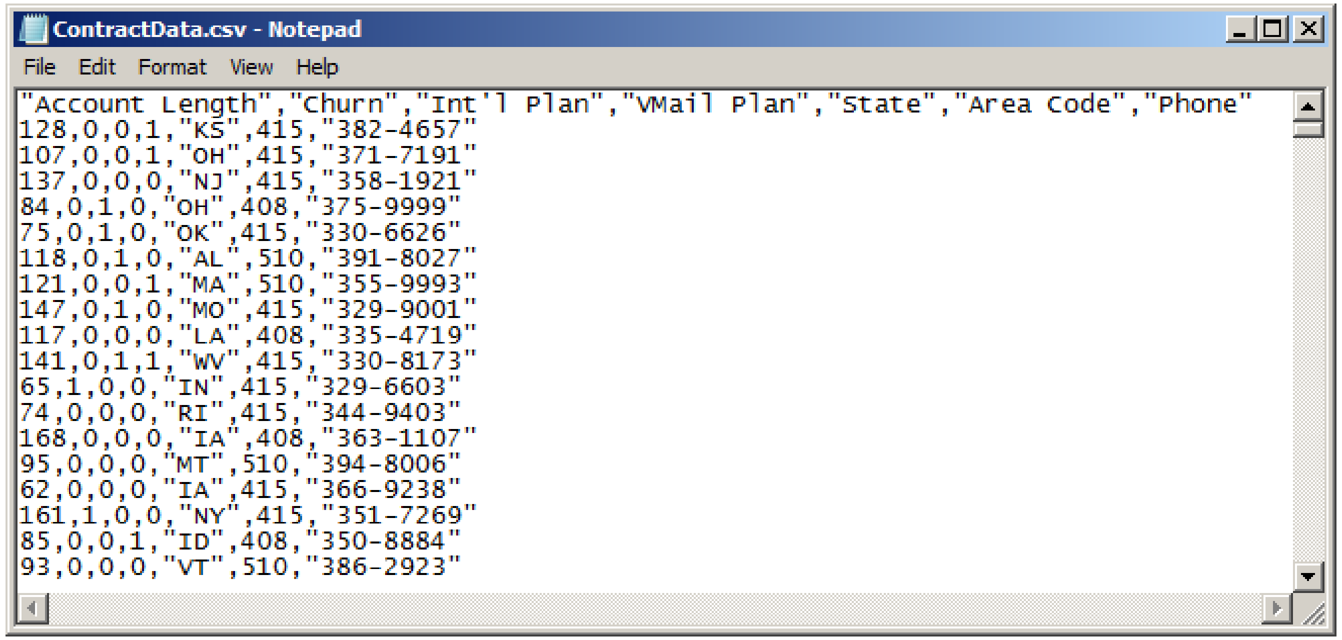
- Data, Dataset, and Scales of Measurement
- Data Collection
- Working with CSV
- Working with Rational Database
- Working with HTML
- Working with XML and JSON
- Working with APIs

Working with CSV Files

Comma Separated Values (CSV) Files

- CSV is a widely used data exchange format. Nowadays many companies are still using CSV to share information with their customers and suppliers.

- CSV files save tabular data in plain text.
  - Use comma (“,”) to separate data fields
  - May use the first line as a header containing field names
  - May use single or double quotation marks around some or all fields



```
"Account Length","Churn","Int'l Plan","VMail Plan","State","Area Code","Phone"
128,0,0,1,"KS",415,"382-4657"
107,0,0,1,"OH",415,"371-7191"
137,0,0,0,"NJ",415,"358-1921"
84,0,1,0,"OH",408,"375-9999"
75,0,1,0,"OK",415,"330-6626"
118,0,1,0,"AL",510,"391-8027"
121,0,0,1,"MA",510,"355-9993"
147,0,1,0,"MO",415,"329-9001"
117,0,0,0,"LA",408,"335-4719"
141,0,1,1,"WV",415,"330-8173"
65,1,0,0,"IN",415,"329-6603"
74,0,0,0,"RI",415,"344-9403"
168,0,0,0,"IA",408,"363-1107"
95,0,0,0,"MT",510,"394-8006"
62,0,0,0,"IA",415,"366-9238"
161,1,0,0,"NY",415,"351-7269"
85,0,0,1,"ID",408,"350-8884"
93,0,0,0,"VT",510,"386-2923"
```

Source: https://en.wikipedia.org/wiki/Comma-separated_values

Comments on CSV Files

- A very simple design of CSV files is to use plain text to store tabular data.
- CSV files can be easily read or written by using Notepad, MS Excel, and other software tools.
- However, a plain text format is very difficult to store complicated data such as long textual data. Important format information may be lost.

R Code: Read and Write CSV Files

```
# Set working directory to the folder which contains the CSV file
setwd("D:\\Cloud\\Dropbox\\Teaching 2016 Fall\\IST 3420\\03.Data Basics")
# Read CSV file
call_data <- read.csv(file = "CallsData.csv", header = TRUE)
head(call_data)
summary(call_data)

# Another way is to use read.table() function
call_data2 <- read.table("CallsData.csv", header = TRUE, sep = ",")
head(call_data2)
summary(call_data2)

# Select Area Code and Phone number and combine them into a new list
phone <- cbind(call_data$Area.Code, as.character(call_data$Phone))
colnames(phone) <- c("AreaCode","PhoneNum")
# Write the phone data into csv file under the working directory
write.csv(phone, file = "Phone.csv")
# Write the phone data into a .txt file, using tab as field separator
write.table(phone, file = "Phone.txt", sep = "\t")

# Show all files under the working directory
dir()
```

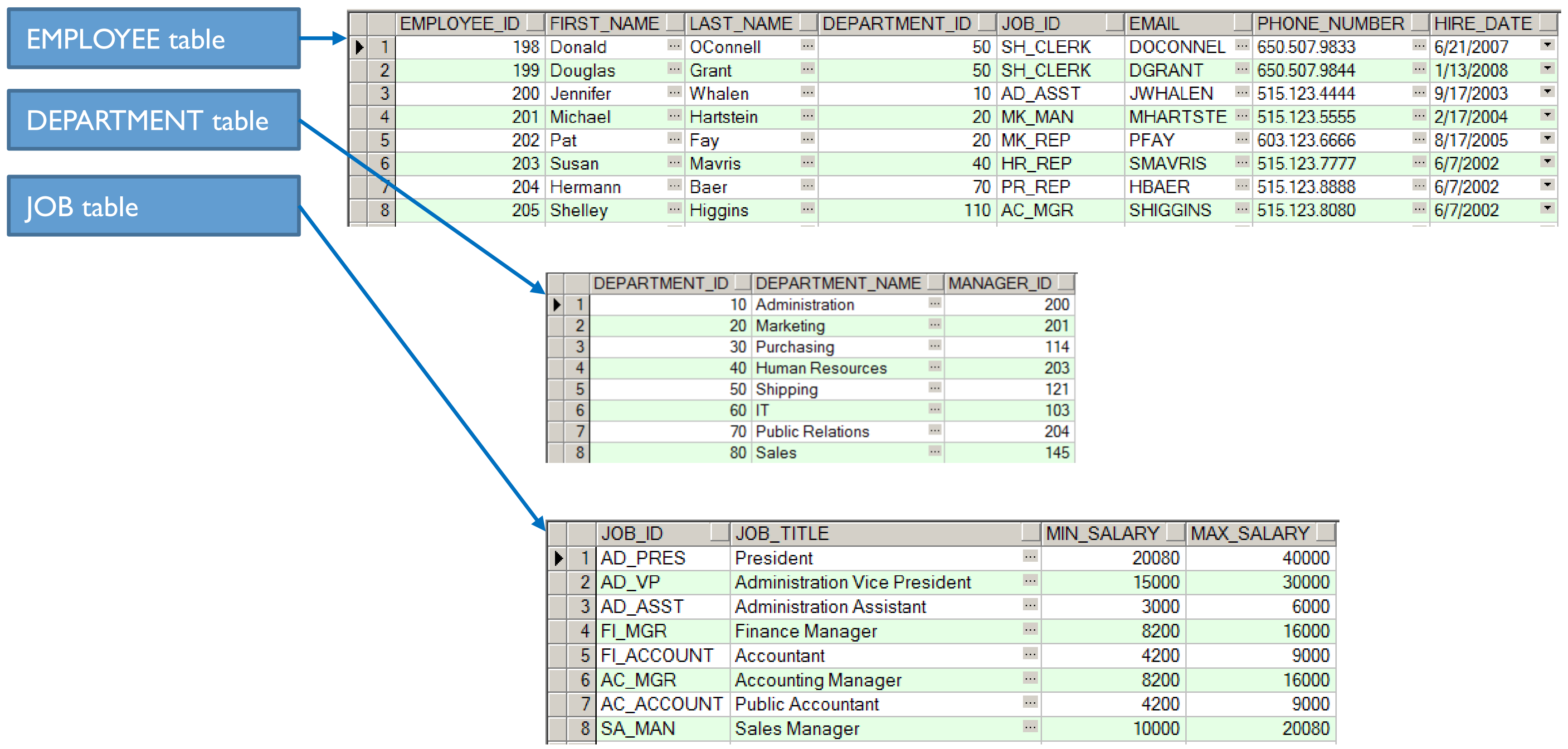
Agenda

- Data, Dataset, and Scales of Measurement
- Data Collection
- Working with CSV
- Working with Rational Database
- Working with HTML
- Working with XML and JSON
- Working with APIs

Working with Relational Database

Relational Database

- A relational database consists of a collection of related data tables



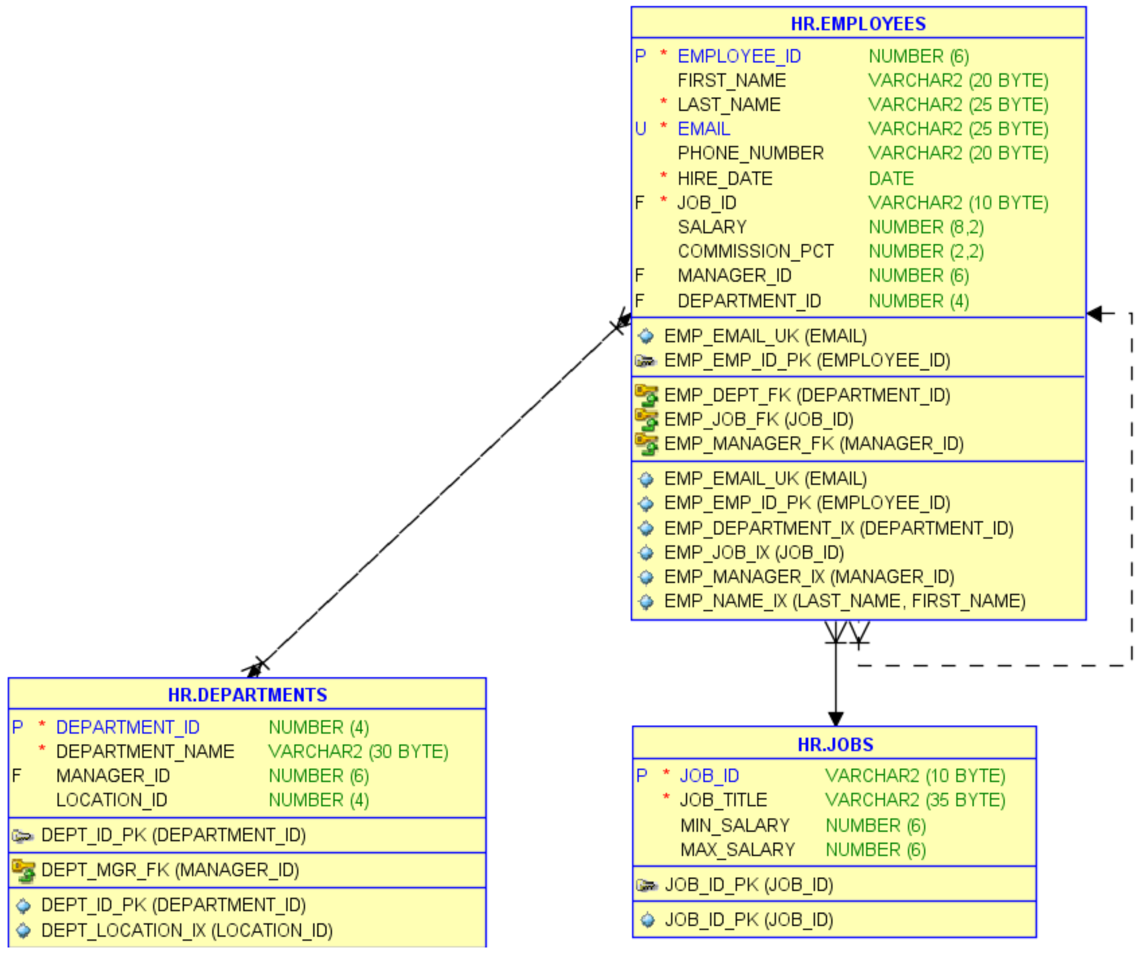
	EMPLOYEE_ID	FIRST_NAME	LAST_NAME	DEPARTMENT_ID	JOB_ID	EMAIL	PHONE_NUMBER	HIRE_DATE
1	198	Donald	OConnell	50	SH_CLERK	DOCONNEL	650.507.9833	6/21/2007
2	199	Douglas	Grant	50	SH_CLERK	DGRANT	650.507.9844	1/13/2008
3	200	Jennifer	Whalen	10	AD_ASST	JWHALEN	515.123.4444	9/17/2003
4	201	Michael	Hartstein	20	MK_MAN	MHARTSTE	515.123.5555	2/17/2004
5	202	Pat	Fay	20	MK_REP	PFAY	603.123.6666	8/17/2005
6	203	Susan	Mavris	40	HR_REP	SMAVRIS	515.123.7777	6/7/2002
7	204	Hermann	Baer	70	PR_REP	HBAER	515.123.8888	6/7/2002
8	205	Shelley	Higgins	110	AC_MGR	SHIGGINS	515.123.8080	6/7/2002

	DEPARTMENT_ID	DEPARTMENT_NAME	MANAGER_ID
1	10	Administration	200
2	20	Marketing	201
3	30	Purchasing	114
4	40	Human Resources	203
5	50	Shipping	121
6	60	IT	103
7	70	Public Relations	204
8	80	Sales	145

	JOB_ID	JOB_TITLE	MIN_SALARY	MAX_SALARY
1	AD_PRES	President	20080	40000
2	AD_VP	Administration Vice President	15000	30000
3	AD_ASST	Administration Assistant	3000	6000
4	FI_MGR	Finance Manager	8200	16000
5	FI_ACCOUNT	Accountant	4200	9000
6	AC_MGR	Accounting Manager	8200	16000
7	AC_ACCOUNT	Public Accountant	4200	9000
8	SA_MAN	Sales Manager	10000	20080

Entity Relationship Diagrams (ERD)

- An ERD describes business entities and their relationships



HR.EMPLOYEES

P	*	EMPLOYEE_ID	NUMBER (6)
		FIRST_NAME	VARCHAR2 (20 BYTE)
	*	LAST_NAME	VARCHAR2 (25 BYTE)
U	*	EMAIL	VARCHAR2 (25 BYTE)
		PHONE_NUMBER	VARCHAR2 (20 BYTE)
	*	HIRE_DATE	DATE
F	*	JOB_ID	VARCHAR2 (10 BYTE)
		SALARY	NUMBER (8,2)
		COMMISSION_PCT	NUMBER (2,2)
F		MANAGER_ID	NUMBER (6)
F		DEPARTMENT_ID	NUMBER (4)

EMP_EMAIL_UK (EMAIL)
EMP_EMP_ID_PK (EMPLOYEE_ID)

EMP_DEPT_FK (DEPARTMENT_ID)
EMP_JOB_FK (JOB_ID)
EMP_MANAGER_FK (MANAGER_ID)

EMP_EMAIL_UK (EMAIL)
EMP_EMP_ID_PK (EMPLOYEE_ID)
EMP_DEPARTMENT_IX (DEPARTMENT_ID)
EMP_JOB_IX (JOB_ID)
EMP_MANAGER_IX (MANAGER_ID)
EMP_NAME_IX (LAST_NAME, FIRST_NAME)

HR.DEPARTMENTS

P	*	DEPARTMENT_ID	NUMBER (4)
	*	DEPARTMENT_NAME	VARCHAR2 (30 BYTE)
F		MANAGER_ID	NUMBER (6)
		LOCATION_ID	NUMBER (4)

DEPT_ID_PK (DEPARTMENT_ID)

DEPT_MGR_FK (MANAGER_ID)

DEPT_ID_PK (DEPARTMENT_ID)
DEPT_LOCATION_IX (LOCATION_ID)

HR.JOBS

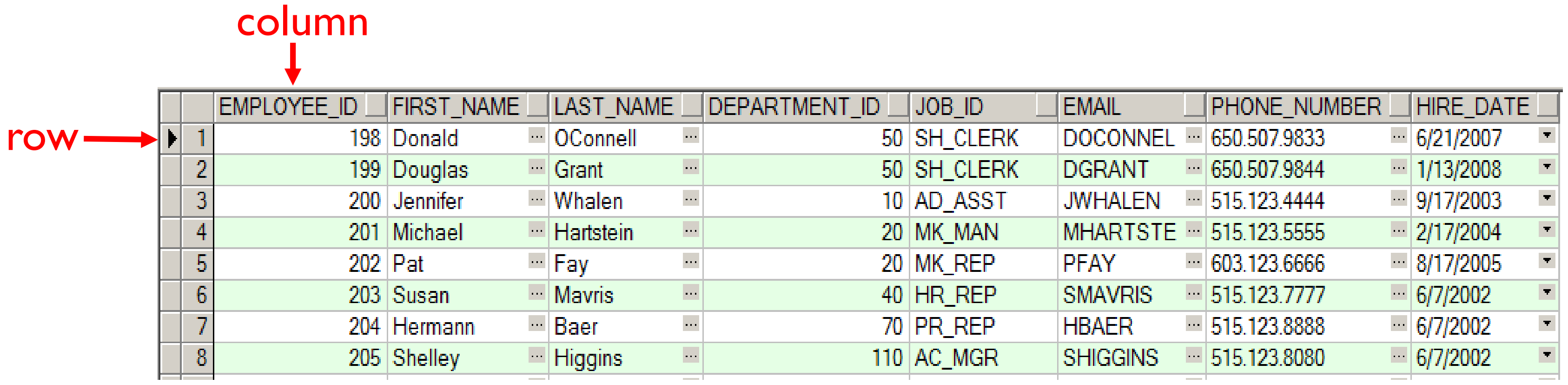
P	*	JOB_ID	VARCHAR2 (10 BYTE)
	*	JOB_TITLE	VARCHAR2 (35 BYTE)
		MIN_SALARY	NUMBER (6)
		MAX_SALARY	NUMBER (6)

JOB_ID_PK (JOB_ID)

JOB_ID_PK (JOB_ID)

Database Tables

- A table is a collection of related data held in a structured format within a database. It consists of columns, and rows.
  - A column (aka attribute, or field) defies the characteristics of the data stored in the column
  - A row (aka record, or tuple) represents a single, implicitly structured data item



	EMPLOYEE_ID	FIRST_NAME	LAST_NAME	DEPARTMENT_ID	JOB_ID	EMAIL	PHONE_NUMBER	HIRE_DATE
1	198	Donald	OConnell	50	SH_CLERK	DOCONNEL	650.507.9833	6/21/2007
2	199	Douglas	Grant	50	SH_CLERK	DGRANT	650.507.9844	1/13/2008
3	200	Jennifer	Whalen	10	AD_ASST	JWHALEN	515.123.4444	9/17/2003
4	201	Michael	Hartstein	20	MK_MAN	MHARTSTE	515.123.5555	2/17/2004
5	202	Pat	Fay	20	MK_REP	PFAY	603.123.6666	8/17/2005
6	203	Susan	Mavris	40	HR_REP	SMAVRIS	515.123.7777	6/7/2002
7	204	Hermann	Baer	70	PR_REP	HBAER	515.123.8888	6/7/2002
8	205	Shelley	Higgins	110	AC_MGR	SHIGGINS	515.123.8080	6/7/2002

Structured Query Language (SQL)

- A query language designed to manipulate data held in a relational DBMS
- Initially developed at IBM in the early 1970s
- Became a standard of the American National Standards Institute (ANSI) in 1986
- Beyond the ANSI-standard SQL, variants are supported by different DBMS platforms
  - Oracle: PL/SQL
  - Microsoft SQL Server: Transact-SQL
  - IBM DB2: SQL PL
  - MySQL: SQL/PSM

Source: https://en.wikipedia.org/wiki/SQL

SQL Select Statement

- SELECT is the most common operation in SQL.
- SELECT retrieves data from one or more tables, or expressions.
- Standard SELECT statements have no persistent effects on the database.

- Syntax:

```
SELECT {ColumnName(s)}
  FROM  {TableName(s)}
  WHERE  {Condition(s)}
```

An Example: Reading Data from SQLite

- Using SQL is a critical skill for data scientists.
- SQLite is a popular embedded RDBMS software for client/local storage.
- Need to install RSQLite package.

Data

- Telecommunication Service Data: Calls and Contracts

Calls Data (TTL 3333 records)

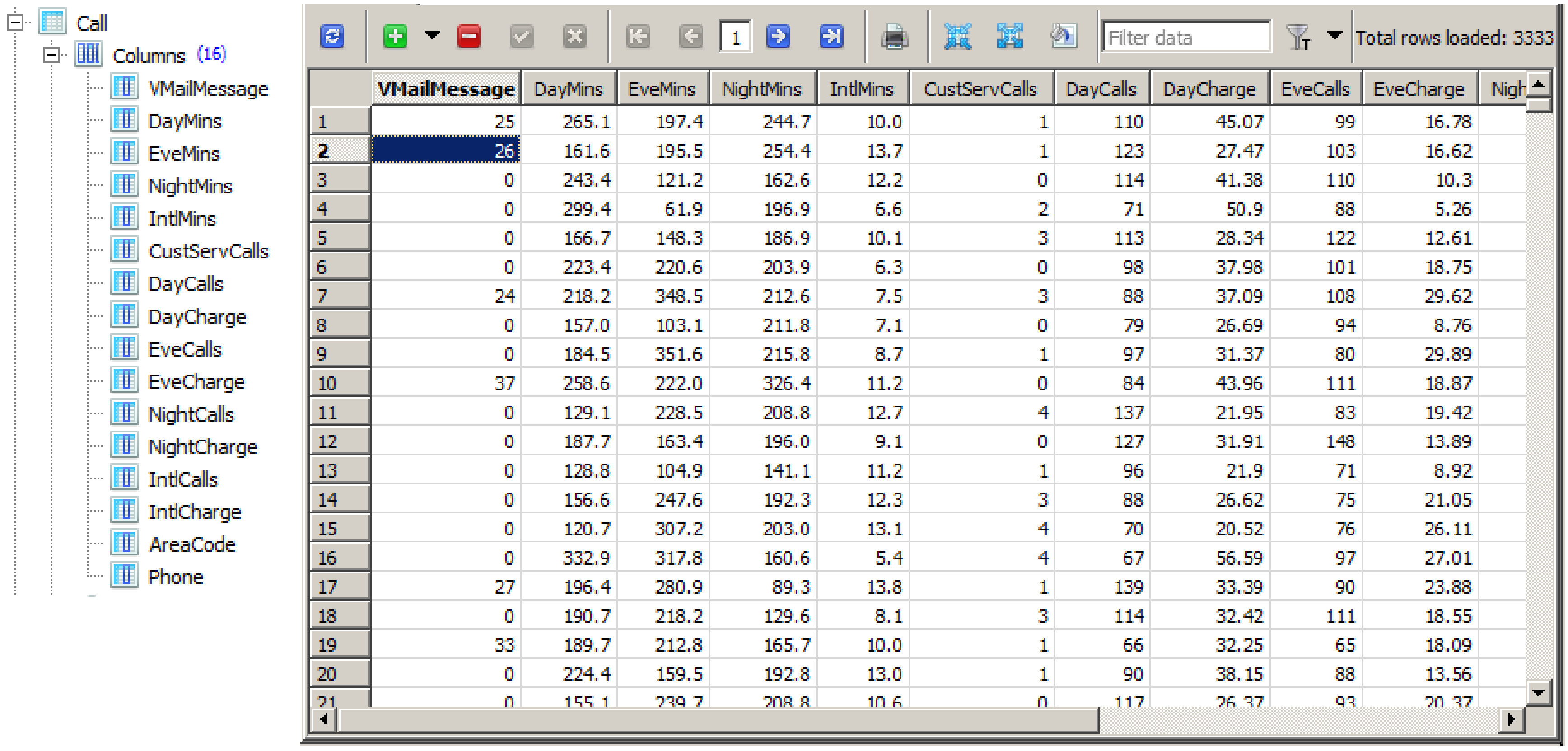
Attribute	Record1	Record2	Record3
VMail Message	25	26	0
Day Mins	265.1	161.6	243.4
Eve Mins	197.4	195.5	121.2
Night Mins	244.7	254.4	162.6
Intl Mins	10	13.7	12.2
CustServ Calls	1	1	0
Day Calls	110	123	114
Day Charge	45.07	27.47	41.38
Eve Calls	99	103	110
Eve Charge	16.78	16.62	10.3
Night Calls	91	103	104
Night Charge	11.01	11.45	7.32
Intl Calls	3	3	5
Intl Charge	2.7	3.7	3.29
Area Code	415	415	415
Phone	382-4657	371-7191	358-1921

Contract Data(TTL 3333 records)

Attribute	Record1	Record2	Record3
Account Length	128	107	137
Churn	0	0	0
Int'l Plan	0	0	0
VMail Plan	1	1	0
State	KS	OH	NJ
Area Code	415	415	415
Phone	382-4657	371-7191	358-1921

SQLite Structure

- Table Call



Call
- Columns (16)
  - VMailMessage
  - DayMins
  - EveMins
  - NightMins
  - IntlMins
  - CustServCalls
  - DayCalls
  - DayCharge
  - EveCalls
  - EveCharge
  - NightCalls
  - NightCharge
  - IntlCalls
  - IntlCharge
  - AreaCode
  - Phone

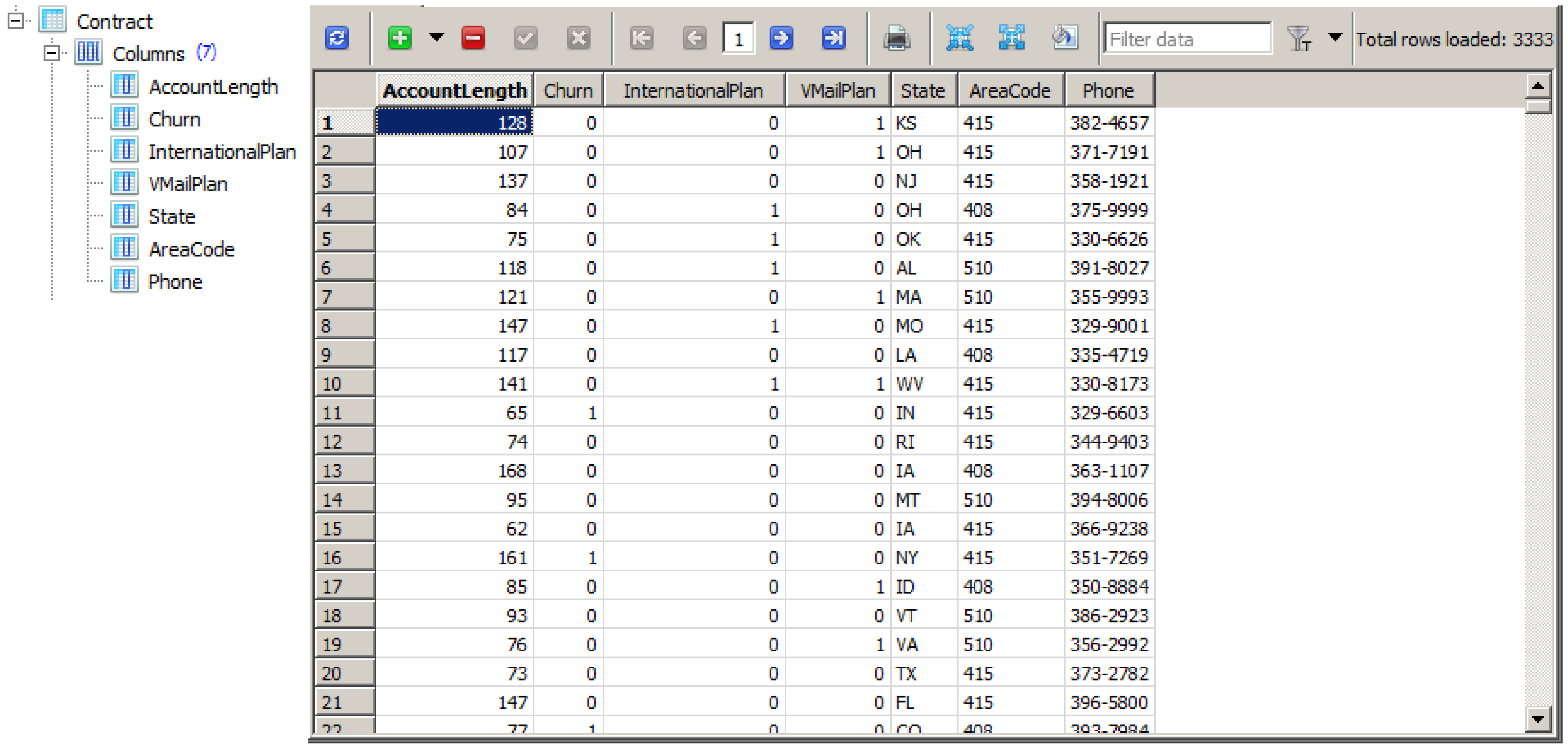
Filter data

Total rows loaded: 3333

	VMailMessage	DayMins	EveMins	NightMins	IntlMins	CustServCalls	DayCalls	DayCharge	EveCalls	EveCharge	Nigh
1	25	265.1	197.4	244.7	10.0	1	110	45.07	99	16.78	
2	26	161.6	195.5	254.4	13.7	1	123	27.47	103	16.62	
3	0	243.4	121.2	162.6	12.2	0	114	41.38	110	10.3	
4	0	299.4	61.9	196.9	6.6	2	71	50.9	88	5.26	
5	0	166.7	148.3	186.9	10.1	3	113	28.34	122	12.61	
6	0	223.4	220.6	203.9	6.3	0	98	37.98	101	18.75	
7	24	218.2	348.5	212.6	7.5	3	88	37.09	108	29.62	
8	0	157.0	103.1	211.8	7.1	0	79	26.69	94	8.76	
9	0	184.5	351.6	215.8	8.7	1	97	31.37	80	29.89	
10	37	258.6	222.0	326.4	11.2	0	84	43.96	111	18.87	
11	0	129.1	228.5	208.8	12.7	4	137	21.95	83	19.42	
12	0	187.7	163.4	196.0	9.1	0	127	31.91	148	13.89	
13	0	128.8	104.9	141.1	11.2	1	96	21.9	71	8.92	
14	0	156.6	247.6	192.3	12.3	3	88	26.62	75	21.05	
15	0	120.7	307.2	203.0	13.1	4	70	20.52	76	26.11	
16	0	332.9	317.8	160.6	5.4	4	67	56.59	97	27.01	
17	27	196.4	280.9	89.3	13.8	1	139	33.39	90	23.88	
18	0	190.7	218.2	129.6	8.1	3	114	32.42	111	18.55	
19	33	189.7	212.8	165.7	10.0	1	66	32.25	65	18.09	
20	0	224.4	159.5	192.8	13.0	1	90	38.15	88	13.56	
21	0	155.1	239.7	208.8	10.6	0	117	26.37	93	20.37	

(cont.)

- Table Contract



Contract
- Columns (7)
  - AccountLength
  - Churn
  - InternationalPlan
  - VMailPlan
  - State
  - AreaCode
  - Phone

	AccountLength	Churn	InternationalPlan	VMailPlan	State	AreaCode	Phone
1	128	0	0	1	KS	415	382-4657
2	107	0	0	1	OH	415	371-7191
3	137	0	0	0	NJ	415	358-1921
4	84	0	1	0	OH	408	375-9999
5	75	0	1	0	OK	415	330-6626
6	118	0	1	0	AL	510	391-8027
7	121	0	0	1	MA	510	355-9993
8	147	0	1	0	MO	415	329-9001
9	117	0	0	0	LA	408	335-4719
10	141	0	1	1	WV	415	330-8173
11	65	1	0	0	IN	415	329-6603
12	74	0	0	0	RI	415	344-9403
13	168	0	0	0	IA	408	363-1107
14	95	0	0	0	MT	510	394-8006
15	62	0	0	0	IA	415	366-9238
16	161	1	0	0	NY	415	351-7269
17	85	0	0	1	ID	408	350-8884
18	93	0	0	0	VT	510	386-2923
19	76	0	0	1	VA	510	356-2992
20	73	0	0	0	TX	415	373-2782
21	147	0	0	0	FL	415	396-5800
22	77	1	0	0	CO	408	393-7984

R Code: Read from SQLite

```
# If you haven't installed RSQLite package, do it.
install.packages("RSQLite")

library("RSQLite")

# Set working directory to the folder which contains the SQLite database file
setwd("D:\\Cloud\\Dropbox\\Teaching 2016 Fall\\IST 3420\\03.Data Basics")

# Specify the database file
dbfile <- "TelecomService.sqlite"

# Create a database connection
con <- dbConnect(dbDriver("SQLite"), dbname = dbfile)

# Get call data
call_data <- dbGetQuery(con, "SELECT * FROM Call")
head(call_data)
summary(call_data)

# Get contract data
contract_data <- dbGetQuery(con, "SELECT * FROM Contract")
head(contract_data)
summary(contract_data)
```

Agenda

- Data, Dataset, and Scales of Measurement
- Data Collection
- Working with CSV
- Working with Rational Database
- Working with HTML
- Working with XML and JSON
- Working with APIs

Working with HTML

HTML Overview

- HyperText Markup Language (HTML) is the standard markup language used to create web pages.

- Web Application = HTML + CSS + JavaScript
  - HTML: Content and structure of web pages
    - Headings
    - Paragraphs
    - Lists
    - Images...
  - CSS (Cascading Style Sheets): Describe the presentation/display of HTML elements
    - Font
    - Color
    - Border...
  - JavaScript: Control the behavior of HTML elements
    - User interaction
    - Button click
    - Dynamic display

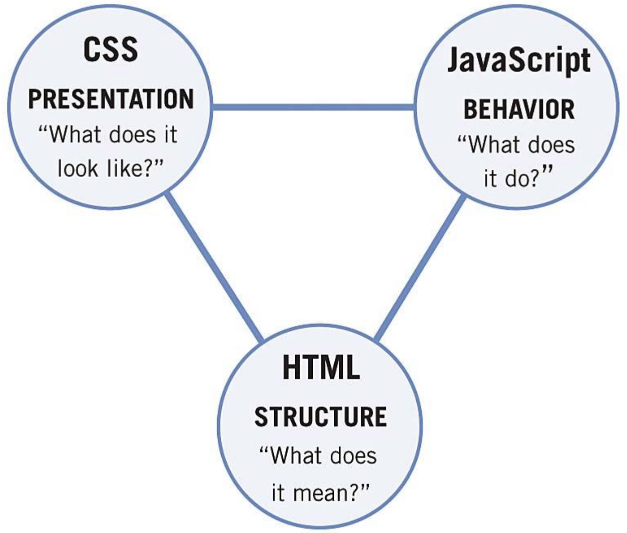


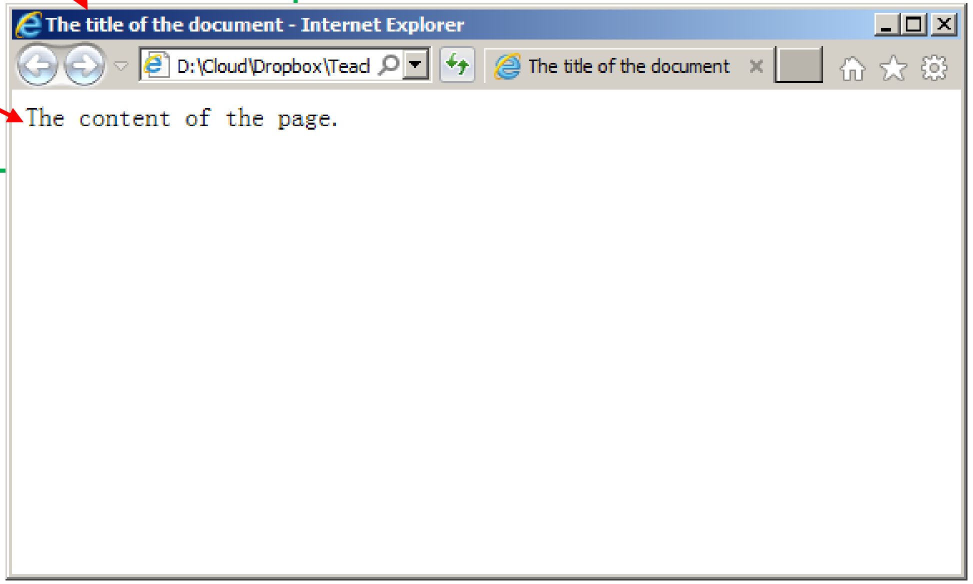
Image source: https://melaloo.files.wordpress.com/2015/07/02fig01.jpg

A Simple HTML Page: A Tree Structure

SimpePage.html

Contents are stored within tags

```
<!DOCTYPE html>
<html>
    <head>
        <title>The title of the document</title>
    </head>
<body>
   The content of the page.
</body>
</html>
```


The title of the document - Internet Explorer
D:\Cloud\Dropbox\Tead
The title of the document

```
The content of the page.
```

URL

- A Uniform Resource Locator (URL) is an identifier which specifies the location of a web resource
  - Web pages (http, https)
  - Email (mailto)
  - FTP
  - Database connection

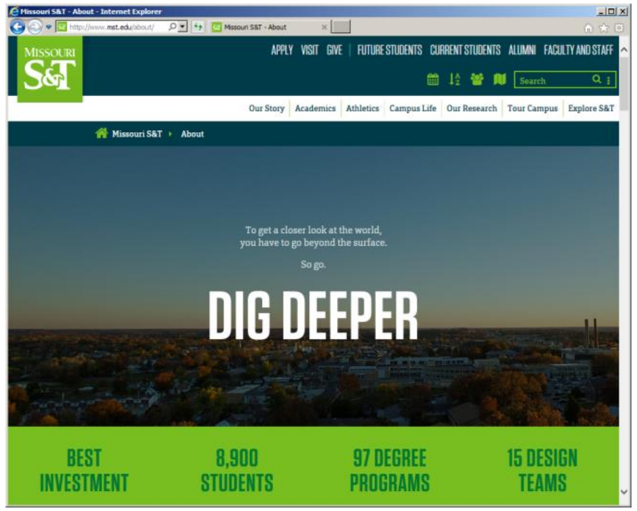
- A web page URL

$$\underbrace{\text{http://}}_{\text{protocol}}\underbrace{\text{www.mst.edu}}_{\text{host name}}/\underbrace{\text{index.html}}_{\text{file location and name}}$$

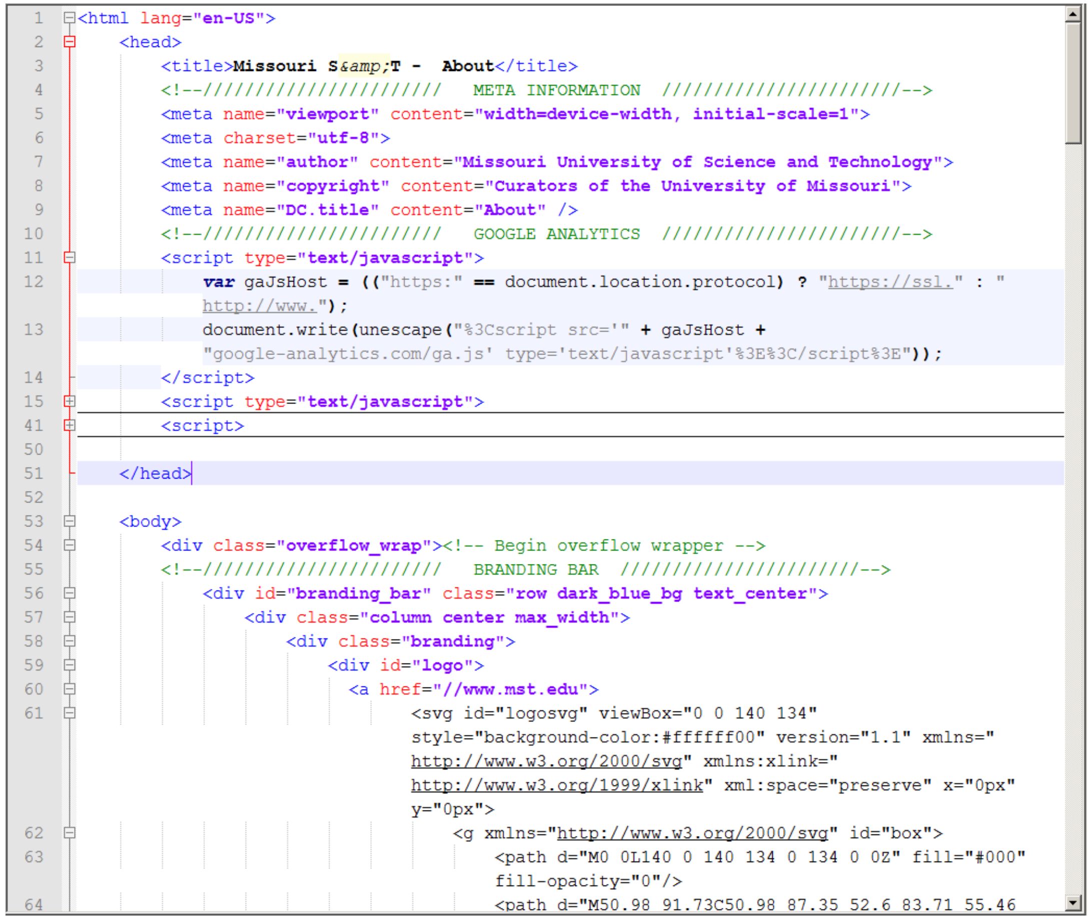
In Practice…

- HTML with CSS and JavaScript becomes complicated

http://www.mst.edu/about/


Missouri S&T - About - Internet Explorer
Missouri S&T
APPLY VISIT GIVE FUTURE STUDENTS CURRENT STUDENTS ALUMNI FACULTY AND STAFF
Search
Our Story Academics Athletics Campus Life Our Research Tour Campus Explore S&T
Missouri S&T › About
DIG DEEPER
BEST INVESTMENT
8,900 STUDENTS
97 DEGREE PROGRAMS
15 DESIGN TEAMS

To get a closer look at the world, you have to go beyond the surface.
So go.



```
<html lang="en-US">
    <head>
        <title>Missouri S&amp;T -  About</title>
        <!--//////////////////////////   META INFORMATION  //////////////////////////-->
        <meta name="viewport" content="width=device-width, initial-scale=1">
        <meta charset="utf-8">
        <meta name="author" content="Missouri University of Science and Technology">
        <meta name="copyright" content="Curators of the University of Missouri">
        <meta name="DC.title" content="About" />
        <!--//////////////////////////   GOOGLE ANALYTICS  //////////////////////////-->
        <script type="text/javascript">
            var gaJsHost = (("https:" == document.location.protocol) ? "https://ssl." : "
            http://www.");
            document.write(unescape("%3Cscript src='" + gaJsHost +
            "google-analytics.com/ga.js' type='text/javascript'%3E%3C/script%3E"));
        </script>
        <script type="text/javascript">
        <script>

    </head>

    <body>
        <div class="overflow_wrap"><!-- Begin overflow wrapper -->
        <!--//////////////////////////   BRANDING BAR  //////////////////////////-->
            <div id="branding_bar" class="row dark_blue_bg text_center">
                <div class="column center max_width">
                    <div class="branding">
                        <div id="logo">
                            <a href="//www.mst.edu">
                                <svg id="logosvg" viewBox="0 0 140 134"
                                style="background-color:#ffffff00" version="1.1" xmlns="
                                http://www.w3.org/2000/svg" xmlns:xlink="
                                http://www.w3.org/1999/xlink" xml:space="preserve" x="0px"
                                y="0px">
                                    <g xmlns="http://www.w3.org/2000/svg" id="box">
                                        <path d="M0 0L140 0 140 134 0 134 0 0Z" fill="#000"
                                        fill-opacity="0"/>
                                        <path d="M50.98 91.73C50.98 87.35 52.6 83.71 55.46
```

Interest in Data Science

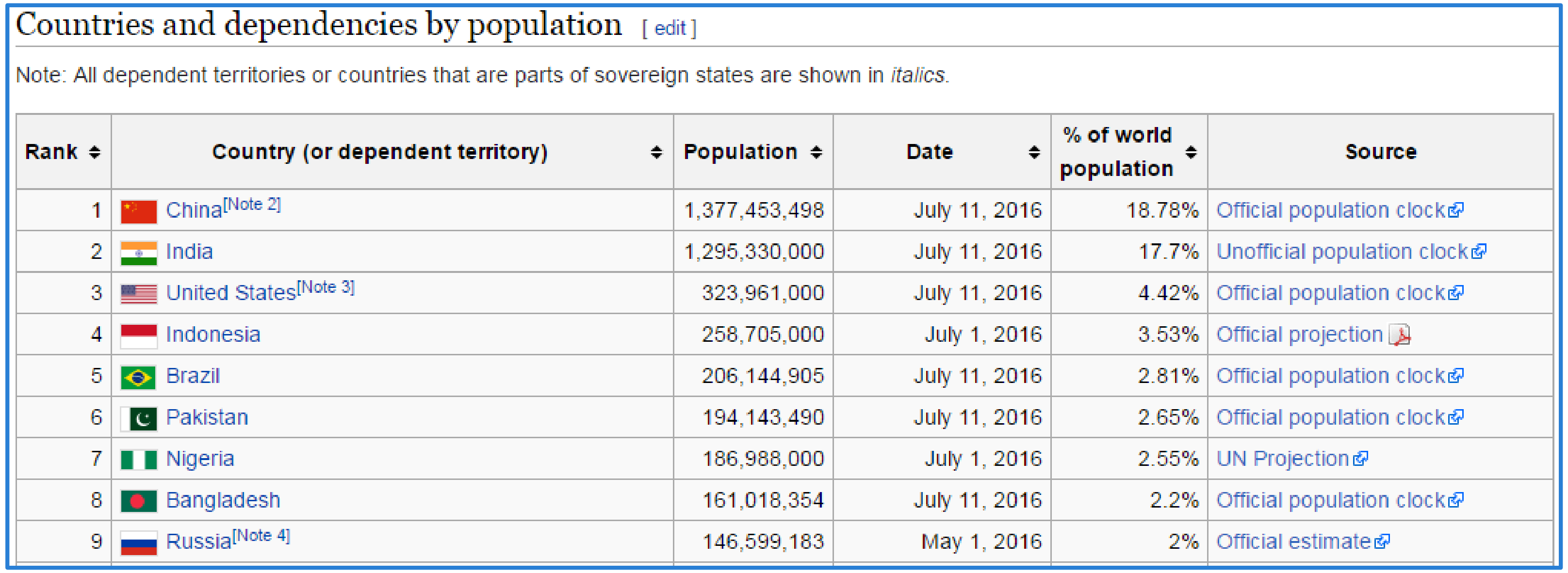
- Our objective is to get data of interest from the complicated web page structure.
- Usually we can ignore all CSS and JavaScript parts.

Guidelines for Web Data Collection

- Inspect the website to fully understand the web page structure and content
- Two important packages in R
  - RCurl
  - XML
- Steps in scraping web pages
  - Specify URL
  - Get the web page content
  - Parse the web page content
  - Read specific elements in the web page

Example: Scraping Tables from Web Pages

- Task: To collect word population data from the following Wikipedia web page https://en.wikipedia.org/wiki/List_of_countries_and_dependencies_by_population



Countries and dependencies by population [edit]

Note: All dependent territories or countries that are parts of sovereign states are shown in *italics*.

Rank	Country (or dependent territory)	Population	Date	% of world population	Source
1	China[Note 2]	1,377,453,498	July 11, 2016	18.78%	Official population clock
2	India	1,295,330,000	July 11, 2016	17.7%	Unofficial population clock
3	United States[Note 3]	323,961,000	July 11, 2016	4.42%	Official population clock
4	Indonesia	258,705,000	July 1, 2016	3.53%	Official projection
5	Brazil	206,144,905	July 11, 2016	2.81%	Official population clock
6	Pakistan	194,143,490	July 11, 2016	2.65%	Official population clock
7	Nigeria	186,988,000	July 1, 2016	2.55%	UN Projection
8	Bangladesh	161,018,354	July 11, 2016	2.2%	Official population clock
9	Russia[Note 4]	146,599,183	May 1, 2016	2%	Official estimate

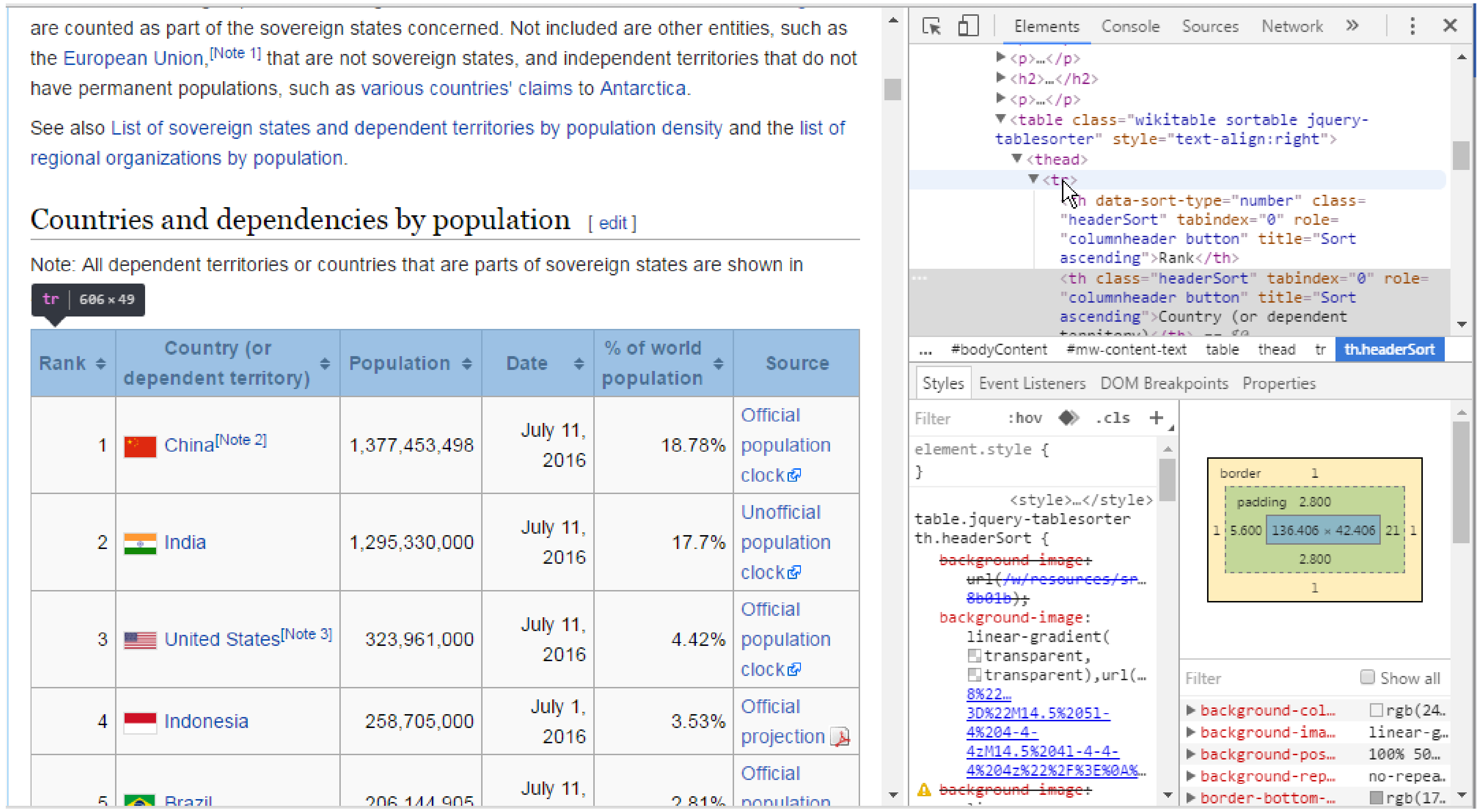
HTML Table Tags

Tag	Description
<table>	Defines a table
<caption>	Defines a table caption
<th>	Defines a header cell in a table
<tr>	Defines a row in a table
<td>	Defines a cell in a table
<thead>	Groups the header content in a table
<tbody>	Groups the body content in a table
<tfoot>	Groups the footer content in a table
<col>	Specifies column properties for each column within a <colgroup> element
<colgroup>	Specifies a group of one or more columns in a table for formatting

For a description of complete HTML tags, refer to: http://www.w3schools.com/tags/ref_byfunc.asp

Inspect the Tree Structure of Web Page

- In Google Chrome, right click the table and then select "Inspect" menu.
- Then move mouse over the HTML element, the corresponding display content will be highlighted.



are counted as part of the sovereign states concerned. Not included are other entities, such as the European Union,[Note 1] that are not sovereign states, and independent territories that do not have permanent populations, such as various countries' claims to Antarctica.

See also List of sovereign states and dependent territories by population density and the list of regional organizations by population.

Countries and dependencies by population [edit]

Note: All dependent territories or countries that are parts of sovereign states are shown in

tr | 606 × 49

Rank	Country (or dependent territory)	Population	Date	% of world population	Source
1	China[Note 2]	1,377,453,498	July 11, 2016	18.78%	Official population clock
2	India	1,295,330,000	July 11, 2016	17.7%	Unofficial population clock
3	United States[Note 3]	323,961,000	July 11, 2016	4.42%	Official population clock
4	Indonesia	258,705,000	July 1, 2016	3.53%	Official projection
5	Brazil	206,144,905	July 11,	2.81%	Official population

Elements Console Sources Network »

```
▶<p>…</p>
▶<h2>…</h2>
▶<p>…</p>
▼<table class="wikitable sortable jquery-
tablesorter" style="text-align:right">
  ▼<thead>
    ▼<tr>
        <th data-sort-type="number" class=
        "headerSort" tabindex="0" role=
        "columnheader button" title="Sort
        ascending">Rank</th>
        <th class="headerSort" tabindex="0" role=
        "columnheader button" title="Sort
        ascending">Country (or dependent
```

… #bodyContent #mw-content-text table thead tr th.headerSort

Styles Event Listeners DOM Breakpoints Properties

Filter :hov .cls +

```
element.style {
}
              <style>…</style>
table.jquery-tablesorter
th.headerSort {
   ~~background-image:
     url(/w/resources/sr…
     8b01b);~~
   background-image:
     linear-gradient(
     transparent,
     transparent),url(…
     8%22…
     3D%22M14.5%2051-
     4%204-4-
     4zM14.5%2041-4-4-
     4%204z%22%2F%3E%0A%…
   ~~background-image:~~
```

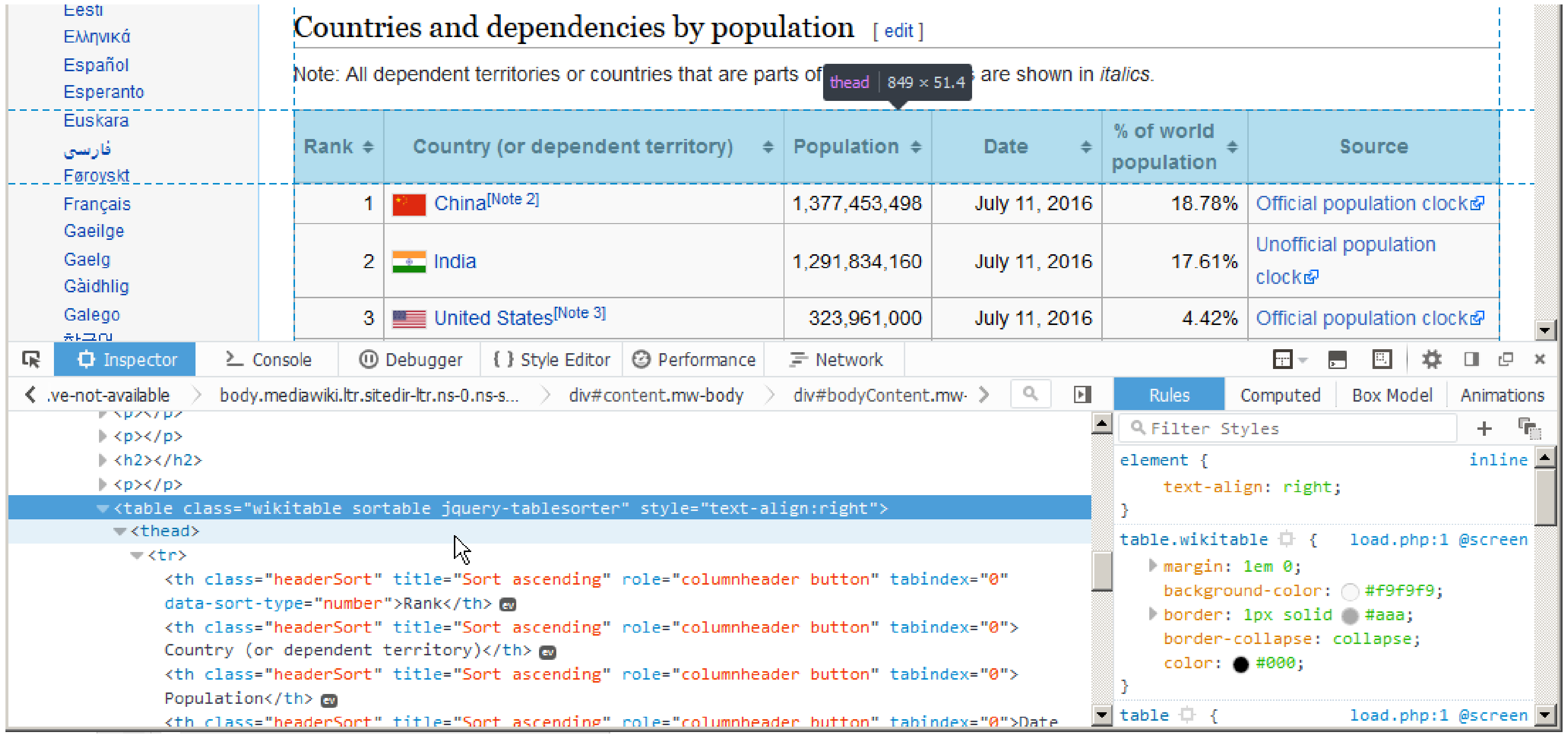
border 1
padding 2.800
1 5.600 136.406 × 42.406 21 1
2.800
1

Filter Show all

```
▶ background-col…   rgb(24…
▶ background-ima…   linear-g…
▶ background-pos…   100% 50…
▶ background-rep…   no-repea…
▶ border-bottom-…   rgb(17…
```

(cont.)

- In Firefox, right click the table and then select "Inspect Element" menu.
- Then move mouse over the HTML element, the corresponding display content will be highlighted.



Countries and dependencies by population [edit]

Note: All dependent territories or countries that are parts of … are shown in *italics*.

Rank	Country (or dependent territory)	Population	Date	% of world population	Source
1	China[Note 2]	1,377,453,498	July 11, 2016	18.78%	Official population clock
2	India	1,291,834,160	July 11, 2016	17.61%	Unofficial population clock
3	United States[Note 3]	323,961,000	July 11, 2016	4.42%	Official population clock

```
<p></p>
<h2></h2>
<p></p>
<table class="wikitable sortable jquery-tablesorter" style="text-align:right">
  <thead>
    <tr>
      <th class="headerSort" title="Sort ascending" role="columnheader button" tabindex="0"
      data-sort-type="number">Rank</th>
      <th class="headerSort" title="Sort ascending" role="columnheader button" tabindex="0">
      Country (or dependent territory)</th>
      <th class="headerSort" title="Sort ascending" role="columnheader button" tabindex="0">
      Population</th>
      <th class="headerSort" title="Sort ascending" role="columnheader button" tabindex="0">Date
```

```
element {                                  inline
    text-align: right;
}
table.wikitable {          load.php:1 @screen
    margin: 1em 0;
    background-color: #f9f9f9;
    border: 1px solid #aaa;
    border-collapse: collapse;
    color: #000;
}
table {                    load.php:1 @screen
```

R Code: Scraping Tables from Web Pages

```
# Load packages
library("XML")
library("RCurl")

# Specify URL
url <- "https://en.wikipedia.org/wiki/List_of_countries_and_dependencies_by_population"
# Download the content of the URL
url_content <- getURL(url)
# Parse the HTML/XML content to generate an R structure representing the HTML/XML tree
doc <- htmlParse(url_content)
tables <- readHTMLTable(doc)

# Convert the 1st element of the list to data frame
pop_df <- data.frame(tables[1])
attributes(pop_df)
colnames(pop_df) <- c("Rank","Country/Territory","Population","Date","% of World Population","Source")

is.factor(pop_df$Population)
# Convert factors into numbers for Population column
pop_df$Population <- as.numeric(gsub(",","",pop_df$Population))
top10 <- head(pop_df, n = 10L)
plot(top10$Population, type = "l")
```

Agenda

- Data, Dataset, and Scales of Measurement
- Data Collection
- Working with CSV
- Working with Rational Database
- Working with HTML
- Working with XML and JSON
- Working with APIs

Working with XML and JSON

What is XML?

- XML stands for EXtensible Markup Language
- XML is a markup language much like HTML
- XML was designed to store and transport data (platform independent)
- XML was designed to be self-descriptive
- XML was designed to be both human-readable and machine-readable
- XML is a W3C Recommendation since 1998

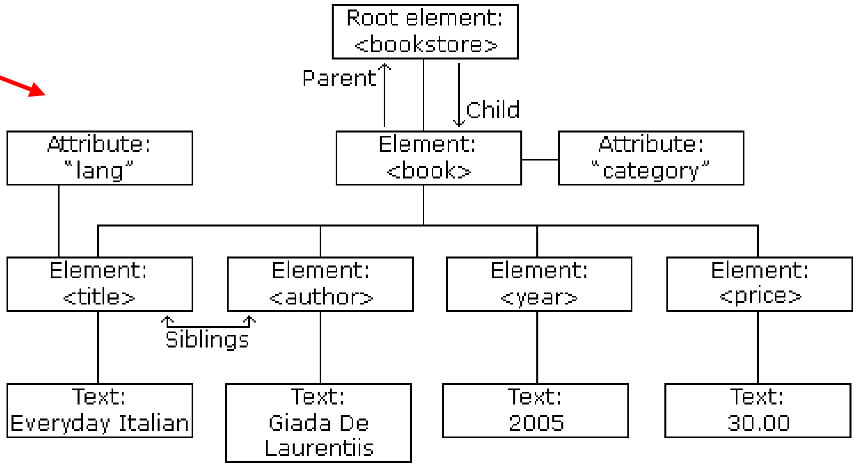
Source: http://www.w3schools.com/xml/xml_whatis.asp

An XML File: Element + Attribute + Text

Books.xml

```
<?xml version="1.0" encoding="UTF-8"?>
<bookstore>

  <book category="cooking">
    <title lang="en">Everyday Italian</title>
    <author>Giada De Laurentiis</author>
    <year>2005</year>
    <price>30.00</price>
  </book>

  <book category="children">
    <title lang="en">Harry Potter</title>
    <author>J K. Rowling</author>
    <year>2005</year>
    <price>29.99</price>
  </book>

  <book category="web" cover="paperback">
    <title lang="en">Learning XML</title>
    <author>Erik T. Ray</author>
    <year>2003</year>
    <price>39.95</price>
  </book>

</bookstore>
```

XML Tree Structure



Source: http://www.w3schools.com/xml/default.asp

Summary of XML Syntax

- An XML document must have a root element
- XML prolog is optional

```
<?xml version="1.0" encoding="UTF-8"?>
```

- Elements must have a closing tag
- Tags are case sensitive
- Elements must be properly nested within each other
- Attributes must be quoted

Source: http://www.w3schools.com/xml/xml_syntax.asp

XML vs. HTML

- XML does not carry any information on data presentation
- You define your own XML tag!
- XML is a complement to HTML
  - Many web applications use XML to store and transport data, while using HTML to display the same data

XPath

- XPath stands for the XML Path Language,
- Use XPath to extract information from XML documents
- XPath uses path expressions to navigate in XML documents

To learn more about XPath, go to:

http://www.w3schools.com/xsl/xpath_intro.asp

Some XPath expressions

XPath Expression	Result
bookstore/book	Selects all book elements that are children of bookstore
/bookstore/book[1]	Selects the first book element that is the child of the bookstore element
/bookstore/book[last()]	Selects the last book element that is the child of the bookstore element
/bookstore/book[position()<3]	Selects the first two book elements that are children of the bookstore element
//@lang	Selects all attributes that are named lang
//title[@lang]	Selects all the title elements that have an attribute named lang
//title[@lang='en']	Selects all the title elements that have a "lang" attribute with a value of "en"
/bookstore/book[price>35.00]	Selects all the book elements of the bookstore element that have a price element with a value greater than 35.00
/bookstore/book[price>35.00]/title	Selects all the title elements of the book elements of the bookstore element that have a price element with a value greater than 35.00

```
<?xml version="1.0" encoding="UTF-
8"?>
<bookstore>

  <book category="cooking">
    <title lang="en">Everyday
Italian</title>
    <author>Giada De
Laurentiis</author>
    <year>2005</year>
    <price>30.00</price>
  </book>

  <book category="children">
    <title lang="en">Harry
Potter</title>
    <author>J K. Rowling</author>
    <year>2005</year>
    <price>29.99</price>
  </book>

  <book category="web"
cover="paperback">
    <title lang="en">Learning
XML</title>
    <author>Erik T. Ray</author>
    <year>2003</year>
    <price>39.95</price>
  </book>

</bookstore>
```

R Code: Read XML Data

```r
require(XML)
# Specify XML file location
loc <- "D:\\Cloud\\Dropbox\\Teaching 2016 Fall\\IST 3420\\3.Data Basics\\Books.xml"

doc <- xmlParse(file = loc) # Parse XML file to generate an R structure
top <- xmlRoot(doc)  # Get the root node

# Explore root node
xmlName(top)
names(top)  # Show child nodes of the root node
names(top[[1]]) # Show child nodes of the 1st book
top[[1]][["title"]] # Show tile of the 1st book
top[[1]][["year"]] # Show year of the 1st book
top[[1]][["price"]] # Show price of the 1st book

# Use xpathSApply() to extract informaiton by XPath expression.
# The xpathSApply() is a simplied version of xpathApply() function.
b_tile <- xpathSApply(doc,"//bookstore/book/title", xmlValue)
b_category <- xpathSApply(doc,"//bookstore/book/@category")
b_author <- xpathSApply(doc,"//bookstore/book/author", xmlValue)
b_year <- xpathSApply(doc,"//bookstore/book/year", xmlValue)
b_price <- xpathSApply(doc,"//bookstore/book/price", xmlValue)

# Generate a book data frame
book_df <- data.frame(b_tile,b_category,b_author,b_year,b_price)
colnames(book_df) <- c("Tile","Category","Author","Year","Price")
print(book_df)
```

What is JSON?

- JSON stands for JavaScript Object Notation, an open-standard format of expressing information.
- A JSON document consists of attribute–value pairs.
- JSON is promoted as a low-overhead alternative to XML.
- Though JSON is named after JavaScript language, it is language independent.
- Many programming languages support JSON.

JSON Syntax

- Data is in name/value pairs (separated by colon)
- Data is separated by commas
- Curly braces hold objects
- Square brackets hold arrays

Source: http://www.w3schools.com/json/

Books.json

```
{
   "Tile":[
      "Everyday Italian",
      "Harry Potter",
      "Learning XML"
   ],
   "Category":[
      "cooking",
      "children",
      "web"
   ],
   "Author":[
      "Giada De Laurentiis",
      "J K. Rowling",
      "Erik T. Ray"
   ],
   "Year":[
      "2005",
      "2005",
      "2003"
   ],
   "Price":[
      "30.00",
      "29.99",
      "39.95"
   ]
}
```

R Code: Access JSON Data

- Use rjson package
  - toJSON() : Convert R to JSON
  - fromJSON() : Convert JSON to R

Note: Use the same data frame object created by XML example

```
# Load rjson package
require(rjson)

# Convert data frame to JSON
# data.frames must be converted into a list before converting into JSON
j_book <- toJSON(as.list(book_df))
print(j_book)

# Convert JSON to data frame
book_df2 <- data.frame(fromJSON(j_book))
print(book_df2)
```

Agenda

- Data, Dataset, and Scales of Measurement
- Data Collection
- Working with CSV
- Working with Rational Database
- Working with HTML
- Working with XML and JSON
- Working with APIs

Working with APIs

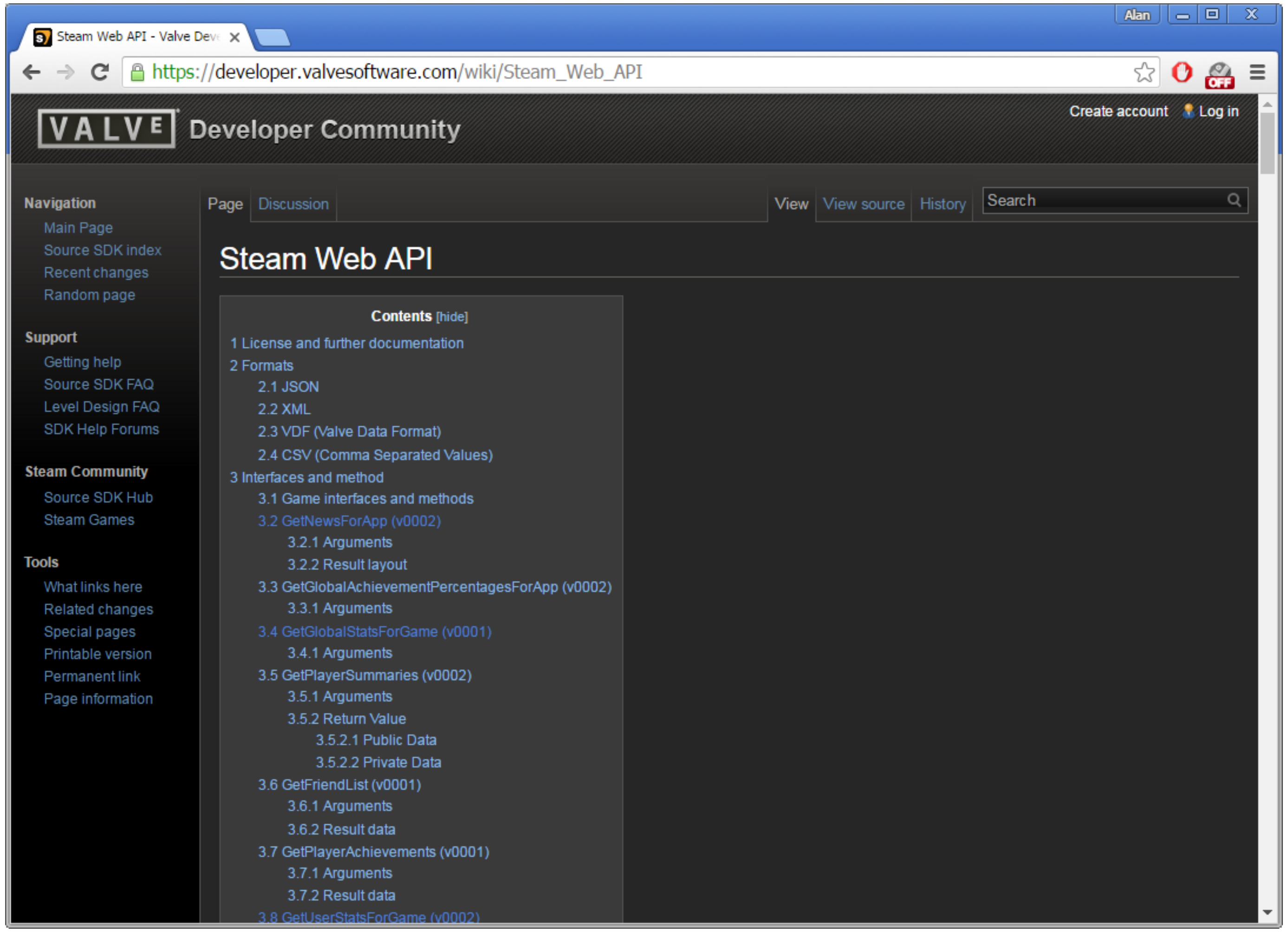
Application Programming Interface (API)

- Some Web Applications provide APIs for access their content.
- APIs usually accept HTTP request.
- We already learn how to access XML and JSON files. Usually the XML and JSON objects are returned by APIs.

- Some APIs
  - YouTube API
  - Google Maps API
  - Twitter API
  - Facebook API
  - Yahoo API
  - LinkedIn API
  - Steam Web API
  - …

Steam Web API Specification

https://developer.valvesoftware.com/wiki/Steam_Web_API



Steam Web API

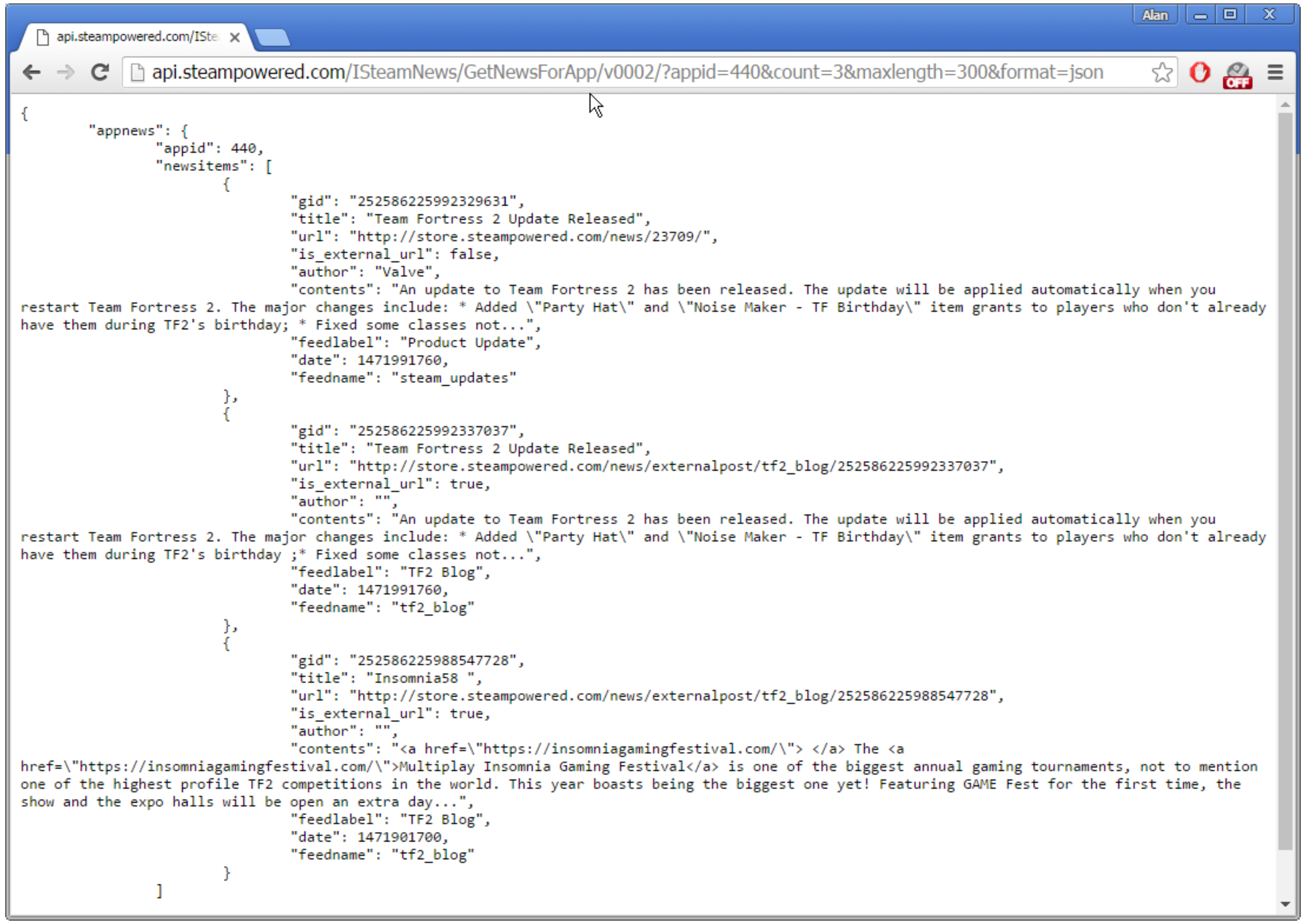
Contents [hide]

GetNewsForApp (v0002)

- GetNewsForApp returns the latest of a game specified by its appID.
- Example URL: http://api.steampowered.com/ISteamNews/GetNewsForApp/v0002/?appid=440&count=3&maxlength=300&format=json
- **Arguments**
  - **appid**
    - AppID of the game you want the news of.
  - **count**
    - How many news entries you want to get returned.
  - **maxlength**
    - Maximum length of each news entry.
  - **format**
    - Output format. *json* (default), *xml* or *vdf*.
- **Result layout**
  - An **appnews** object containing:
  - **appid**, the AppID of the game you want news of
  - **newsitems**, an array of news item information:
    - An ID, title and url.
    - A shortened excerpt of the contents (to maxlength characters), terminated by "..." if longer than maxlength.
    - A comma-separated string of labels and UNIX timestamp.

You can test the API in browsers

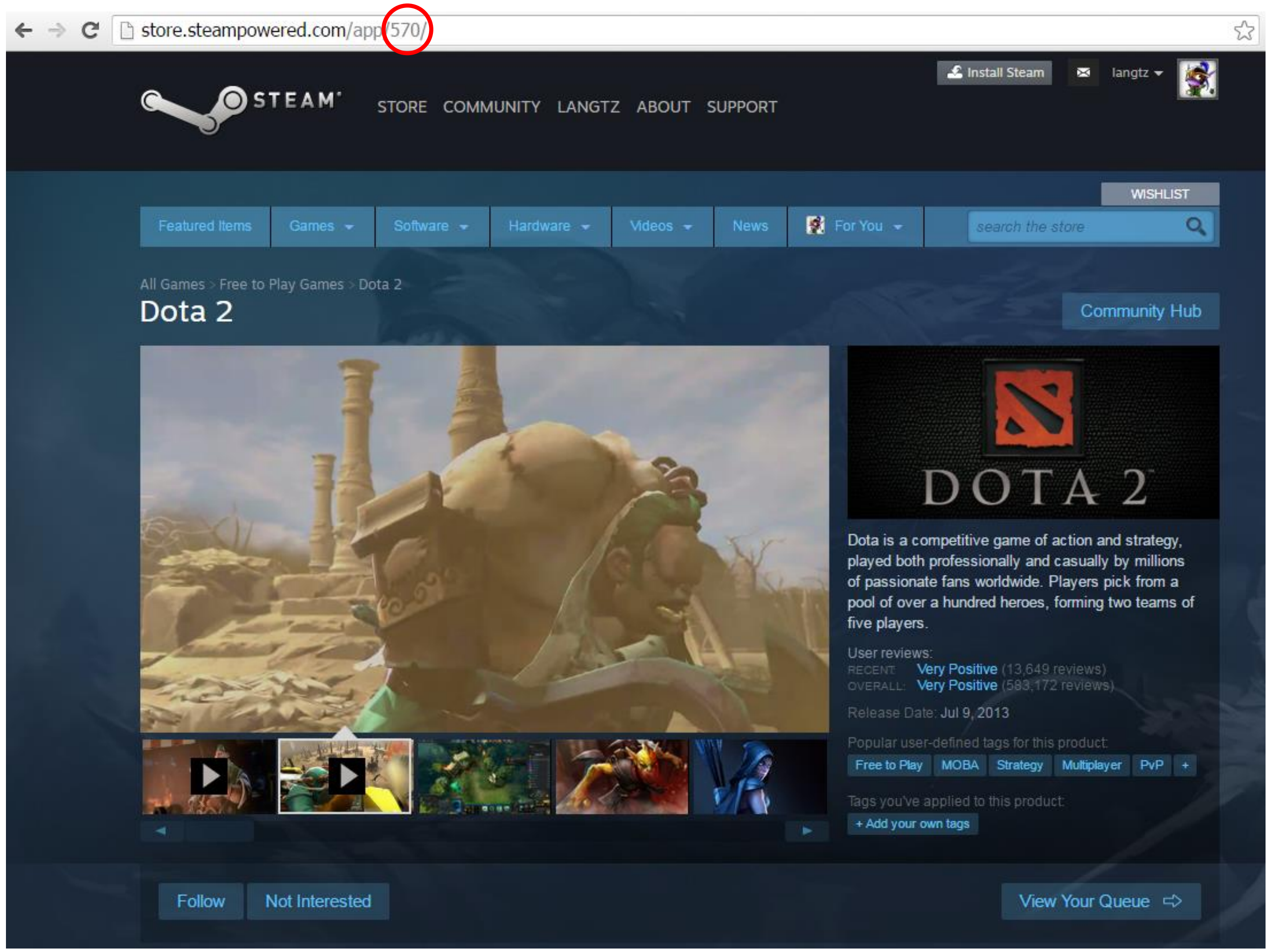
http://api.steampowered.com/ISteamNews/GetNewsForApp/v0002/?appid=440&count=3&maxlength=300&format=json


api.steampowered.com/ISte
Alan
api.steampowered.com/ISteamNews/GetNewsForApp/v0002/?appid=440&count=3&maxlength=300&format=json
OFF

```
{
        "appnews": {
                "appid": 440,
                "newsitems": [
                        {
                                "gid": "252586225992329631",
                                "title": "Team Fortress 2 Update Released",
                                "url": "http://store.steampowered.com/news/23709/",
                                "is_external_url": false,
                                "author": "Valve",
                                "contents": "An update to Team Fortress 2 has been released. The update will be applied automatically when you
restart Team Fortress 2. The major changes include: * Added \"Party Hat\" and \"Noise Maker - TF Birthday\" item grants to players who don't already
have them during TF2's birthday; * Fixed some classes not...",
                                "feedlabel": "Product Update",
                                "date": 1471991760,
                                "feedname": "steam_updates"
                        },
                        {
                                "gid": "252586225992337037",
                                "title": "Team Fortress 2 Update Released",
                                "url": "http://store.steampowered.com/news/externalpost/tf2_blog/252586225992337037",
                                "is_external_url": true,
                                "author": "",
                                "contents": "An update to Team Fortress 2 has been released. The update will be applied automatically when you
restart Team Fortress 2. The major changes include: * Added \"Party Hat\" and \"Noise Maker - TF Birthday\" item grants to players who don't already
have them during TF2's birthday ;* Fixed some classes not...",
                                "feedlabel": "TF2 Blog",
                                "date": 1471991760,
                                "feedname": "tf2_blog"
                        },
                        {
                                "gid": "252586225988547728",
                                "title": "Insomnia58 ",
                                "url": "http://store.steampowered.com/news/externalpost/tf2_blog/252586225988547728",
                                "is_external_url": true,
                                "author": "",
                                "contents": "<a href=\"https://insomniagamingfestival.com/\"> </a> The <a
href=\"https://insomniagamingfestival.com/\">Multiplay Insomnia Gaming Festival</a> is one of the biggest annual gaming tournaments, not to mention
one of the highest profile TF2 competitions in the world. This year boasts being the biggest one yet! Featuring GAME Fest for the first time, the
show and the expo halls will be open an extra day...",
                                "feedlabel": "TF2 Blog",
                                "date": 1471901700,
                                "feedname": "tf2_blog"
                        }
                ]
```

What is an App ID?

- Select an App from steam store (http://store.steampowered.com/) to check it's App ID

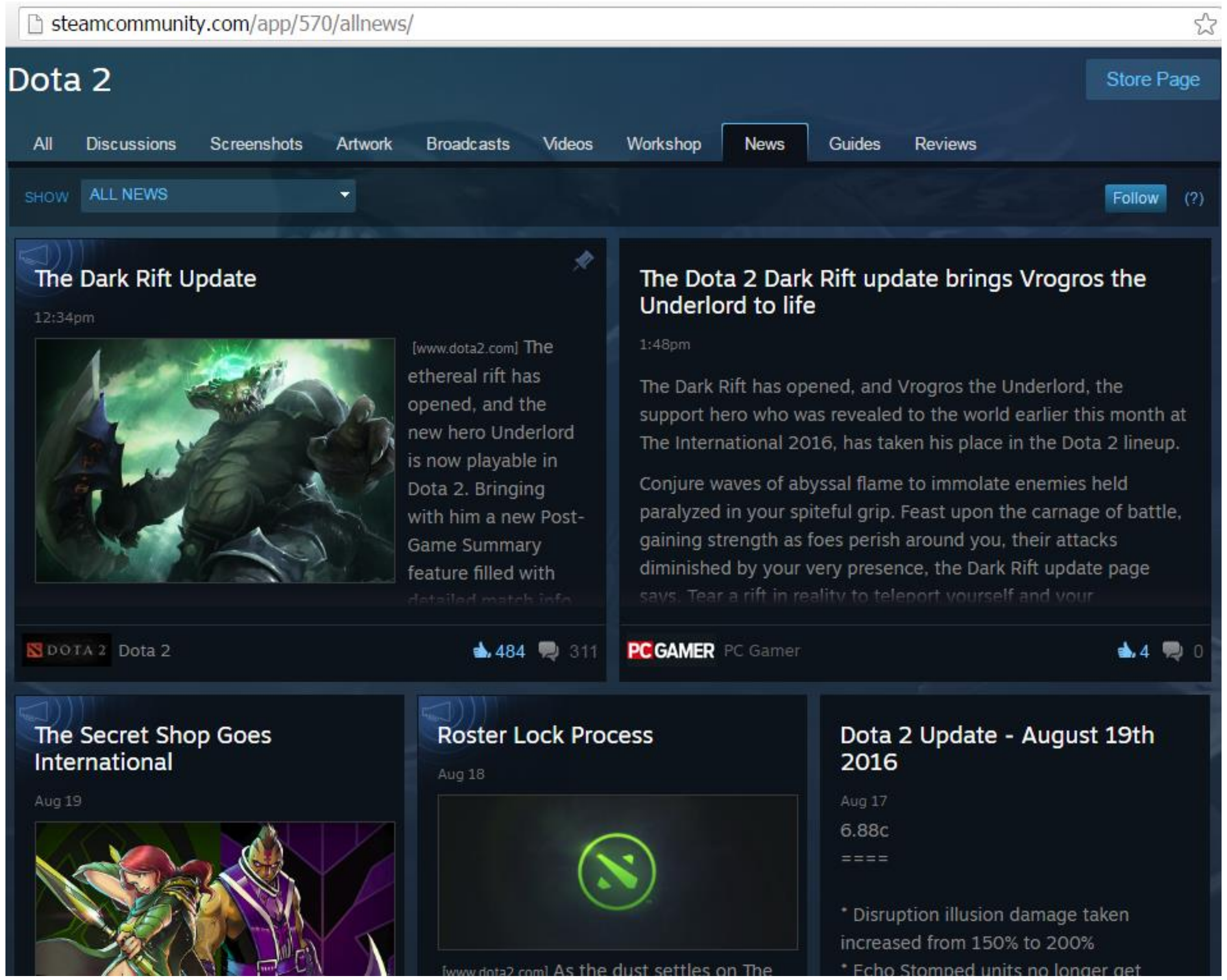


Dota is a competitive game of action and strategy, played both professionally and casually by millions of passionate fans worldwide. Players pick from a pool of over a hundred heroes, forming two teams of five players.

User reviews:
RECENT: Very Positive (13,649 reviews)
OVERALL: Very Positive (583,172 reviews)
Release Date: Jul 9, 2013
Popular user-defined tags for this product:
Tags you've applied to this product:

What data the API provides?

http://steamcommunity.com/app/570/allnews/



The Dark Rift Update

12:34pm

[www.dota2.com] The ethereal rift has opened, and the new hero Underlord is now playable in Dota 2. Bringing with him a new Post-Game Summary feature filled with

Dota 2 484 311

The Dota 2 Dark Rift update brings Vrogros the Underlord to life

1:48pm

The Dark Rift has opened, and Vrogros the Underlord, the support hero who was revealed to the world earlier this month at The International 2016, has taken his place in the Dota 2 lineup.

Conjure waves of abyssal flame to immolate enemies held paralyzed in your spiteful grip. Feast upon the carnage of battle, gaining strength as foes perish around you, their attacks diminished by your very presence, the Dark Rift update page

PC GAMER PC Gamer 4 0

The Secret Shop Goes International

Aug 19

Roster Lock Process

Aug 18

Dota 2 Update - August 19th 2016

Aug 17

6.88c

====

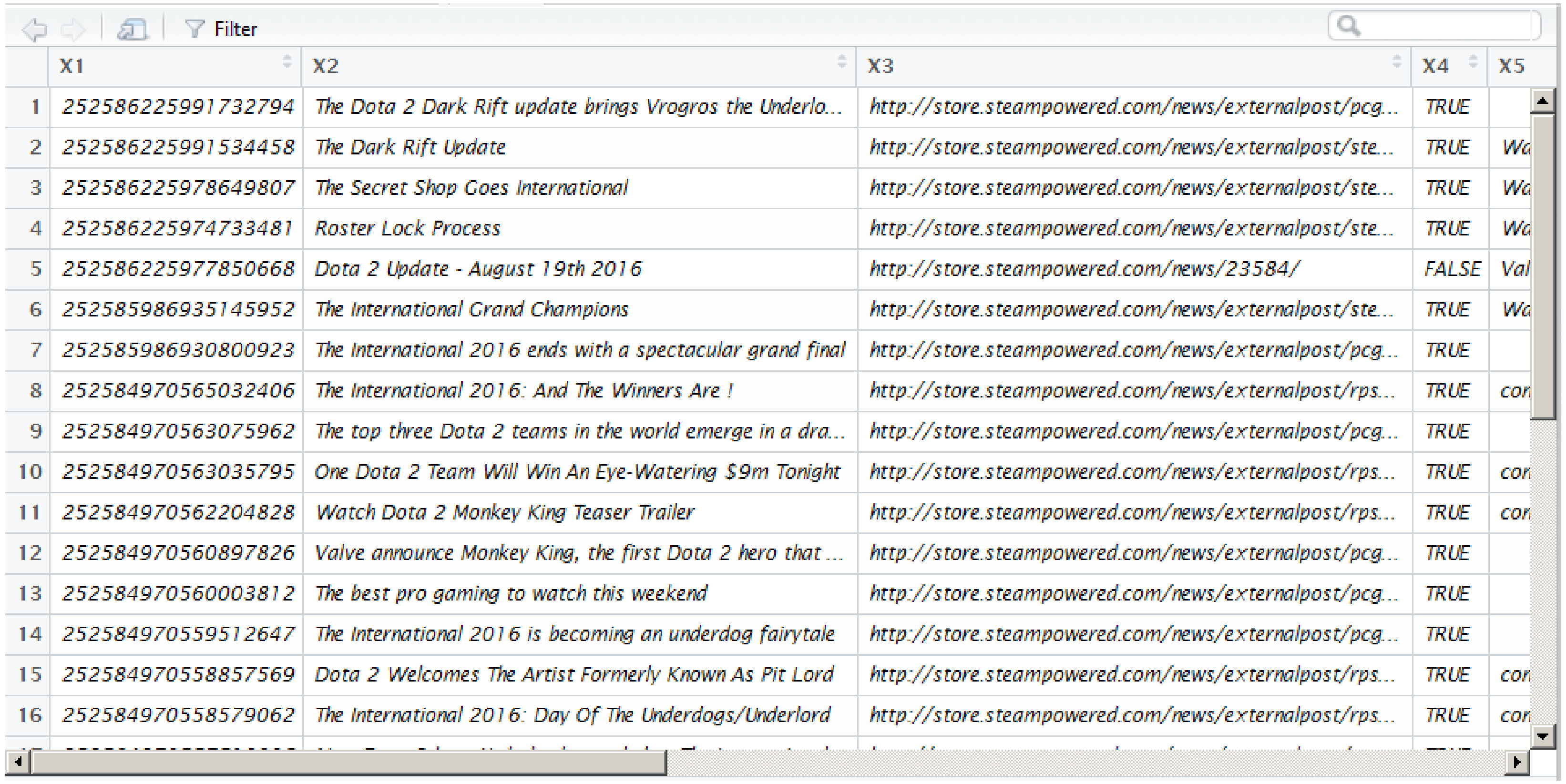
* Disruption illusion damage taken increased from 150% to 200%

R Code

```
appid <- 570 # Choose Dota 2 whose App ID is 570
n_new <- 30 # Number of news to get
# Generate url for the app
url <- paste("http://api.steampowered.com/ISteamNews/GetNewsForApp/v0002/?appid=", appid,
"&count=", n_new, "&maxlength=300&format=json", sep = "")

# Download HTTP response
content = RCurl::getURI(url)

# Create a list
list <- fromJSON(content)
# Show content in the list
list$appnews$newsitems
list$appnews$newsitems[[1]]

# Create a data frame
df <- data.frame(matrix(unlist(list$appnews$newsitems), nrow = n_new, byrow=T))
# Rename variables in the data frame
colnames(df) <- names(json$appnews$newsitems[[1]])
# Show content of the data frame. You'll notice that the date is shown as numbers (in strings actually)
df
# Add variable date2 to show the readable date time
df$date2 <- as.POSIXlt(as.numeric(as.character(df$date)),origin = "1970-01-01")
df
```

Dataset Collected

Data collected on 8/23/2016

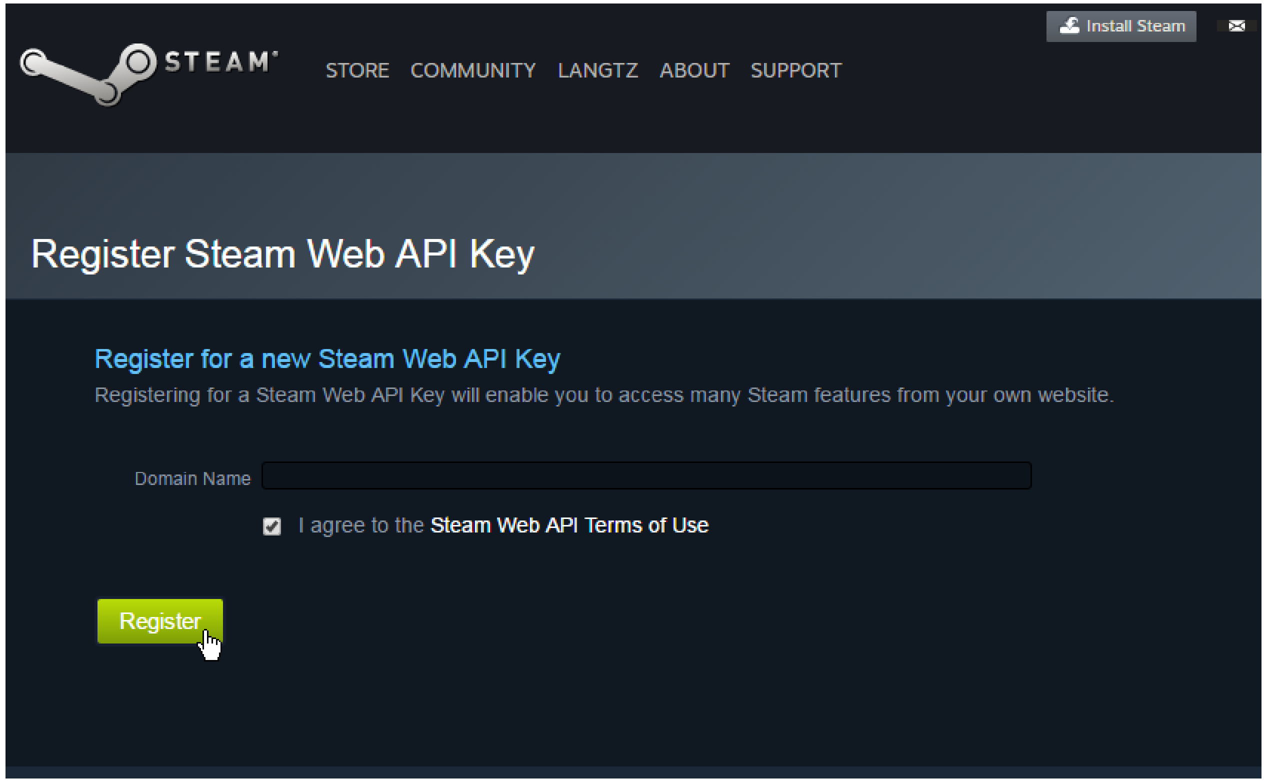


Filter

	X1	X2	X3	X4	X5
1	252586225991732794	The Dota 2 Dark Rift update brings Vrogros the Underlo...	http://store.steampowered.com/news/externalpost/pcg...	TRUE	
2	252586225991534458	The Dark Rift Update	http://store.steampowered.com/news/externalpost/ste...	TRUE	Wa
3	252586225978649807	The Secret Shop Goes International	http://store.steampowered.com/news/externalpost/ste...	TRUE	Wa
4	252586225974733481	Roster Lock Process	http://store.steampowered.com/news/externalpost/ste...	TRUE	Wa
5	252586225977850668	Dota 2 Update - August 19th 2016	http://store.steampowered.com/news/23584/	FALSE	Val
6	252585986935145952	The International Grand Champions	http://store.steampowered.com/news/externalpost/ste...	TRUE	Wa
7	252585986930800923	The International 2016 ends with a spectacular grand final	http://store.steampowered.com/news/externalpost/pcg...	TRUE	
8	252584970565032406	The International 2016: And The Winners Are !	http://store.steampowered.com/news/externalpost/rps...	TRUE	cor
9	252584970563075962	The top three Dota 2 teams in the world emerge in a dra...	http://store.steampowered.com/news/externalpost/pcg...	TRUE	
10	252584970563035795	One Dota 2 Team Will Win An Eye-Watering $9m Tonight	http://store.steampowered.com/news/externalpost/rps...	TRUE	cor
11	252584970562204828	Watch Dota 2 Monkey King Teaser Trailer	http://store.steampowered.com/news/externalpost/rps...	TRUE	cor
12	252584970560897826	Valve announce Monkey King, the first Dota 2 hero that ...	http://store.steampowered.com/news/externalpost/pcg...	TRUE	
13	252584970560003812	The best pro gaming to watch this weekend	http://store.steampowered.com/news/externalpost/pcg...	TRUE	
14	252584970559512647	The International 2016 is becoming an underdog fairytale	http://store.steampowered.com/news/externalpost/pcg...	TRUE	
15	252584970558857569	Dota 2 Welcomes The Artist Formerly Known As Pit Lord	http://store.steampowered.com/news/externalpost/rps...	TRUE	cor
16	252584970558579062	The International 2016: Day Of The Underdogs/Underlord	http://store.steampowered.com/news/externalpost/rps...	TRUE	cor

Some APIs need a Steam Web API Key

- Click the link (http://steamcommunity.com/dev/apikey) to register one



Register Steam Web API Key

Register for a new Steam Web API Key

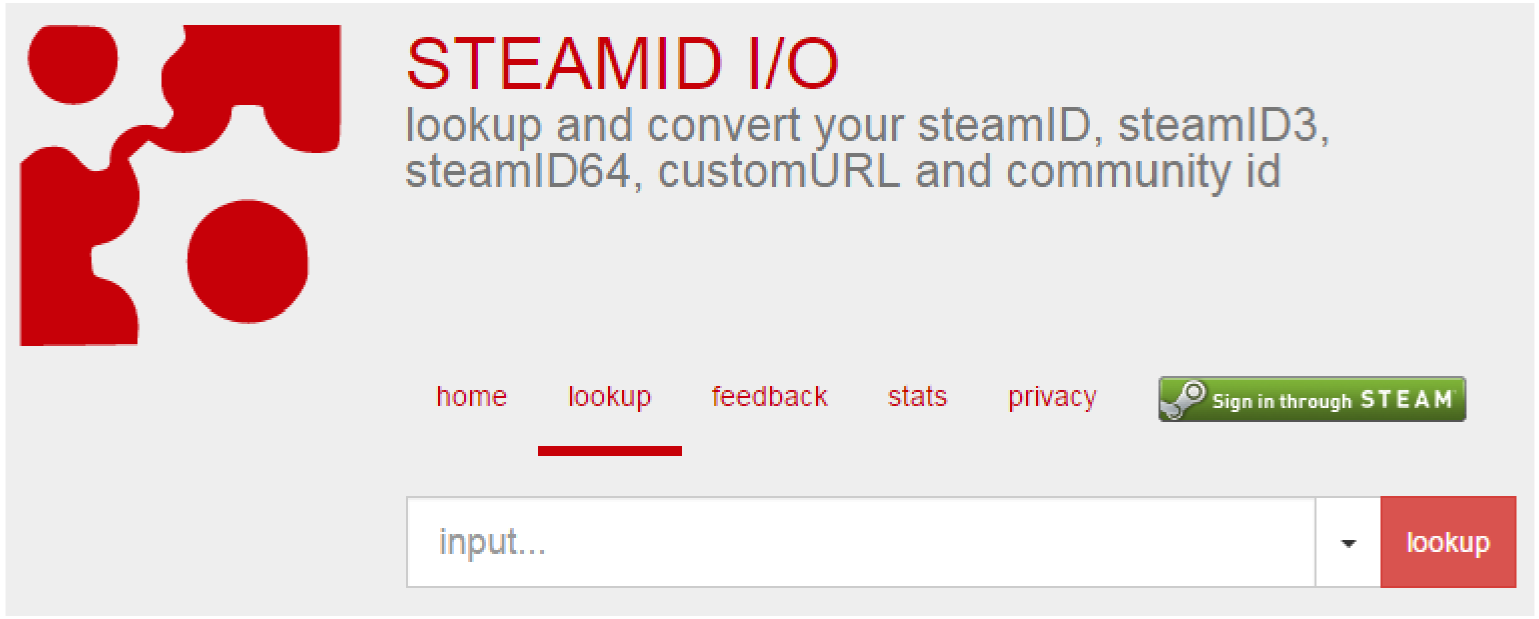
Registering for a Steam Web API Key will enable you to access many Steam features from your own website.

Domain Name

✓ I agree to the Steam Web API Terms of Use

Some APIs need a 64bit SteamID of the user

- Click the link (https://steamid.io/lookup) to check


STEAMID I/O

lookup and convert your steamID, steamID3, steamID64, customURL and community id

Reading Assignment (due Sep 17)

- *Base R Cheat Sheet*
  - https://www.rstudio.com/wp-content/uploads/2016/09/r-cheat-sheet-1.pdf

- *Data Wrangling with dplyr and tidyr Cheat Sheet*
  - https://www.rstudio.com/wp-content/uploads/2015/02/data-wrangling-cheatsheet.pdf

Q & A