



# PREDICT 420

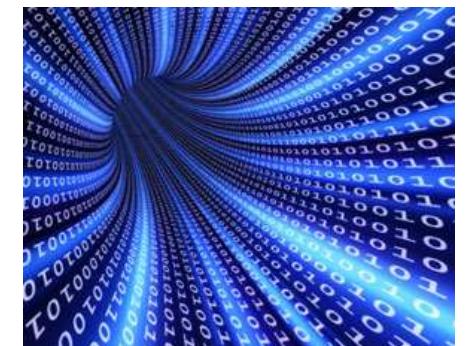
Atef Bader, PhD

# Agenda

---

---

- Syllabus
- Canvas - Course Homepage
- FAQ
- Course Topics - Walkthrough
- Python/Canopy tool
- Exercise #1 Practice
- What you need to submit for Exercise #1 on Canvas?



# Syllabus

PREDICT 420 Syllabus - Summer 2017.pdf - Adobe Acrobat Reader DC

File Edit View Window Help

Home Tools PREDICT 420 Sylla... x Atef

1 / 21 100% 100%

NORTHWESTERN UNIVERSITY SCHOOL OF PROFESSIONAL STUDIES

PREDICT 420: Database Systems and Data Preparation Summer 2017

Instructor Name  
Atef Bader, PhD  
[a-bader@northwestern.edu](mailto:a-bader@northwestern.edu)

Course Description  
Behind every analytics project is an analytical data source. In this course, students explore the fundamentals of data management and data preparation. Students acquire hands-on experience with various data file formats, working with quantitative data and text, relational (SQL) database systems, and NoSQL database systems. They access, organize, clean, prepare, transform, and explore data, using database shells, query and scripting languages, and analytical software. This is a case-study- and project-based course with a strong programming component.

Prerequisites  
PREDICT 400: Math for Modelers for students entering the MSPA program after fall 2014.

Learning Goals

- Articulate analytics as a core strategy using examples of successful predictive modeling/data mining applications in various industries.
- Formulate and manage plans to address business issues with analytics.
- Define key terms, concepts and issues in data management and database management systems with respect to predictive modeling
- Evaluate the constraints, limitations and structure of data through data cleansing, preparation and exploratory analysis to create an analytical database.
- Use object-oriented scripting software for data preparation.

# Canvas Course Homepage Layout

The screenshot shows the Canvas Course Homepage Layout for the course "2017SU\_PREDICT\_420-DL\_SEC56".

**Header:** File Edit View History Bookmarks Tools Help  
2017SU\_PREDICT\_420-DL\_SEC56 X +  
https://canvas.northwestern.edu/courses/57663 Search  
Most Visited Getting Started

**Sidebar (Left):**

- N
- Account
- Dashboard
- Courses
- Calendar
- Inbox
- Help

**Course Navigation:** 2017SU\_PREDICT\_420-DL\_SEC56 > Modules

**Course Information:** 2017 Summer

**Home:** Home, Announcements, Modules, Course Reserves, People, Grades, Syllabus, Library Media, Sync Session, Blue Jeans, CTEC, Library Resources.

**Content Area:**

- PREDICT 420 Database Systems & Data Preparation
  - Getting Started
  - Introductions
  - Course Overview
- Course Resources
  - Software and Systems
    - Python
    - Jupyter Notebook
    - Jupyter notebook - Documentation

**Right Sidebar:**

- View Course Stream
- Coming Up: View Calendar (Nothing for the next week)

**Bottom Bar:**

- You are currently logged into Student View
- Reset Student
- Leave Student View

# FAQ

---

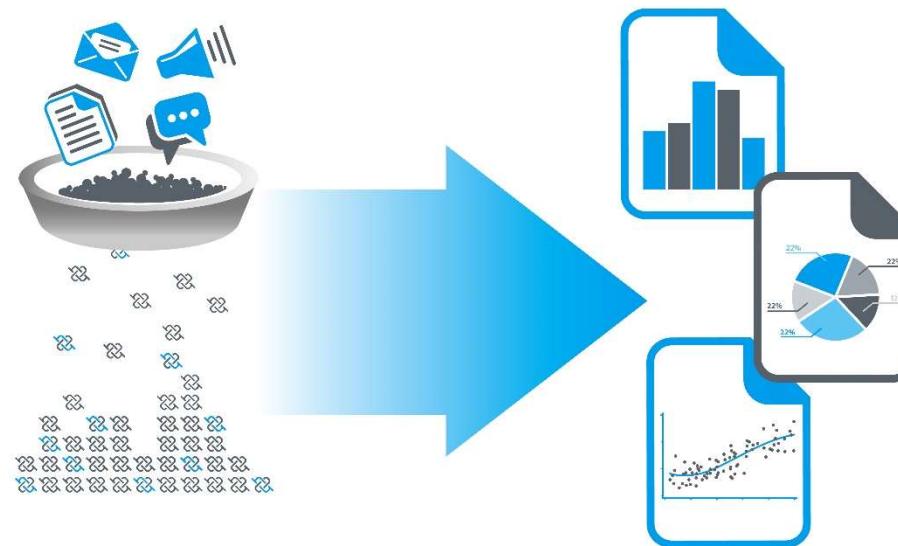
---

- What is this class about?
  - Explore the fundamentals of data management and data preparation
  - Structured data vs. Unstructured data
  - Relational Database & File Processing
  - Python, SQL/NoSQL, Postgres
- Where I can find the weekly recordings for the sync sessions?
  - All sync sessions are recorded on weekly basis and posted on Canvas by 11:59pm On Mondays
- Where can I find the due dates for assignments?
  - Canvas/Syllabus

# Data

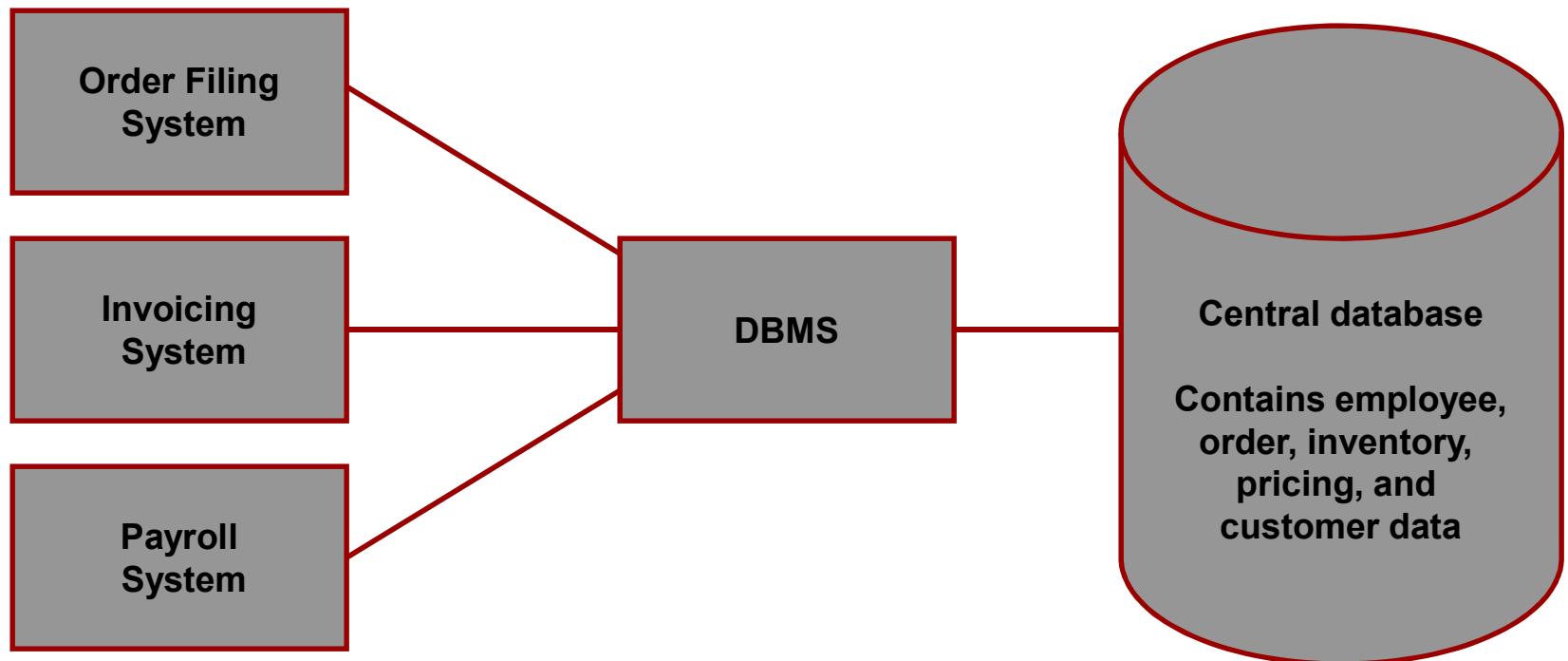
---

- Data could be
  - 1. Structured
    - Employee record, Product, Order, Transaction, etc.
  - 2. Unstructured
    - Email, Tweets, blogs, social chats, reviews, etc.



# Database Management System

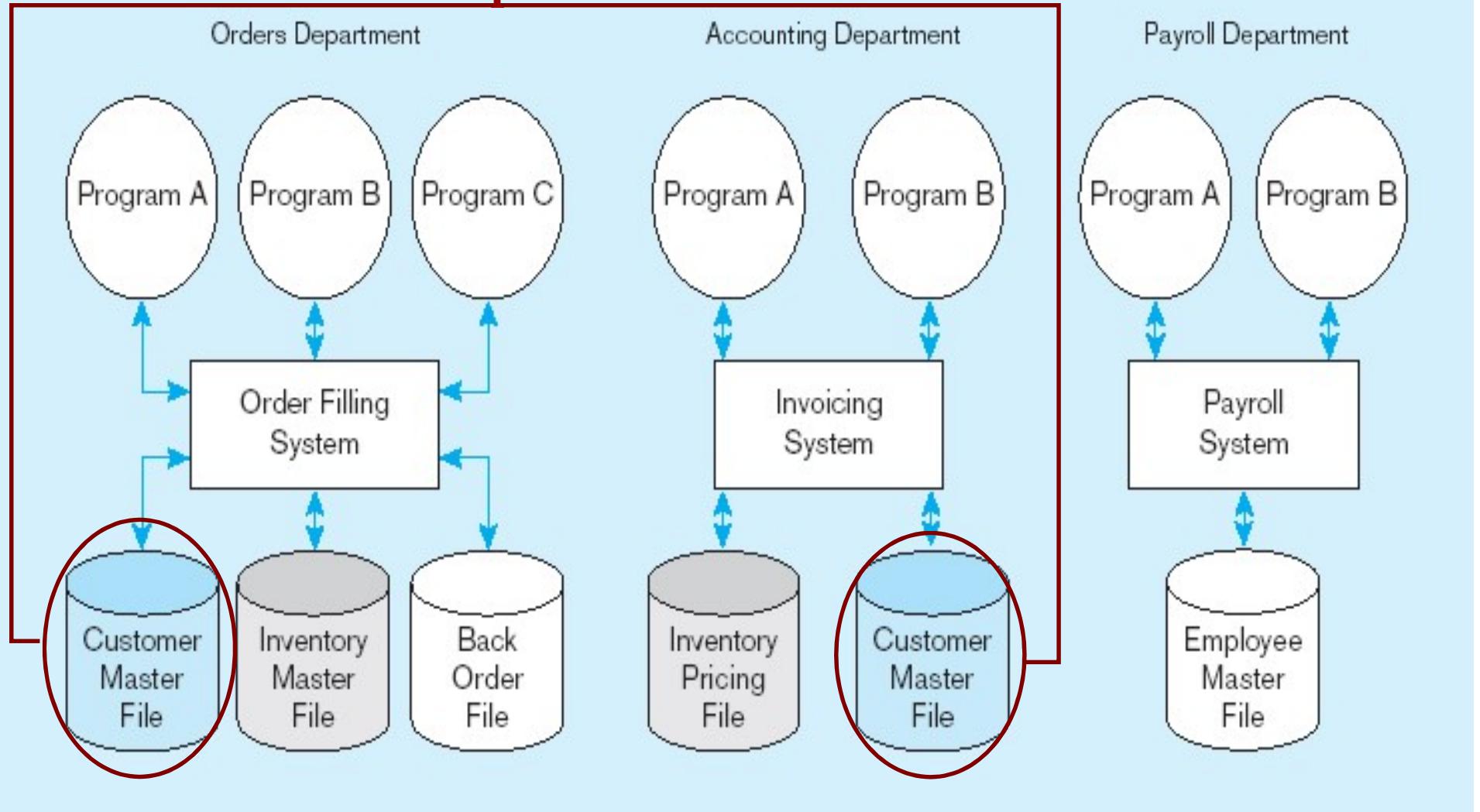
A software system that is used to create, maintain, and provide controlled access to user databases



*DBMS manages data resources like an operating system manages hardware resources*

# File Processing Systems

## Duplicate Data



# **Disadvantages of File Processing**

---

---

## **□ Program-Data Dependence**

- All programs maintain metadata for each file they use

## **□ Duplication of Data**

- Different systems/programs have separate copies of the same data

## **□ Limited Data Sharing**

- No centralized control of data

## **□ Lengthy Development Times**

- Programmers must design their own file formats

## **□ Excessive Program Maintenance**

- 80% of information systems budget

# Advantages of the Database Approach

---

- Program-data independence
- Planned data redundancy
- Improved data consistency
- Improved data sharing
- Increased application development productivity
- Enforcement of standards
- Improved data quality
- Improved data accessibility and responsiveness
- Reduced program maintenance
- Improved decision support

# **Old School for Data vs New School for Data**

---

## **□ The Old School for Data**

- Unstructured Data (Megabytes/Gigabytes)
- Structured Data → Relations (Tables)
- File Processing → RDBMS



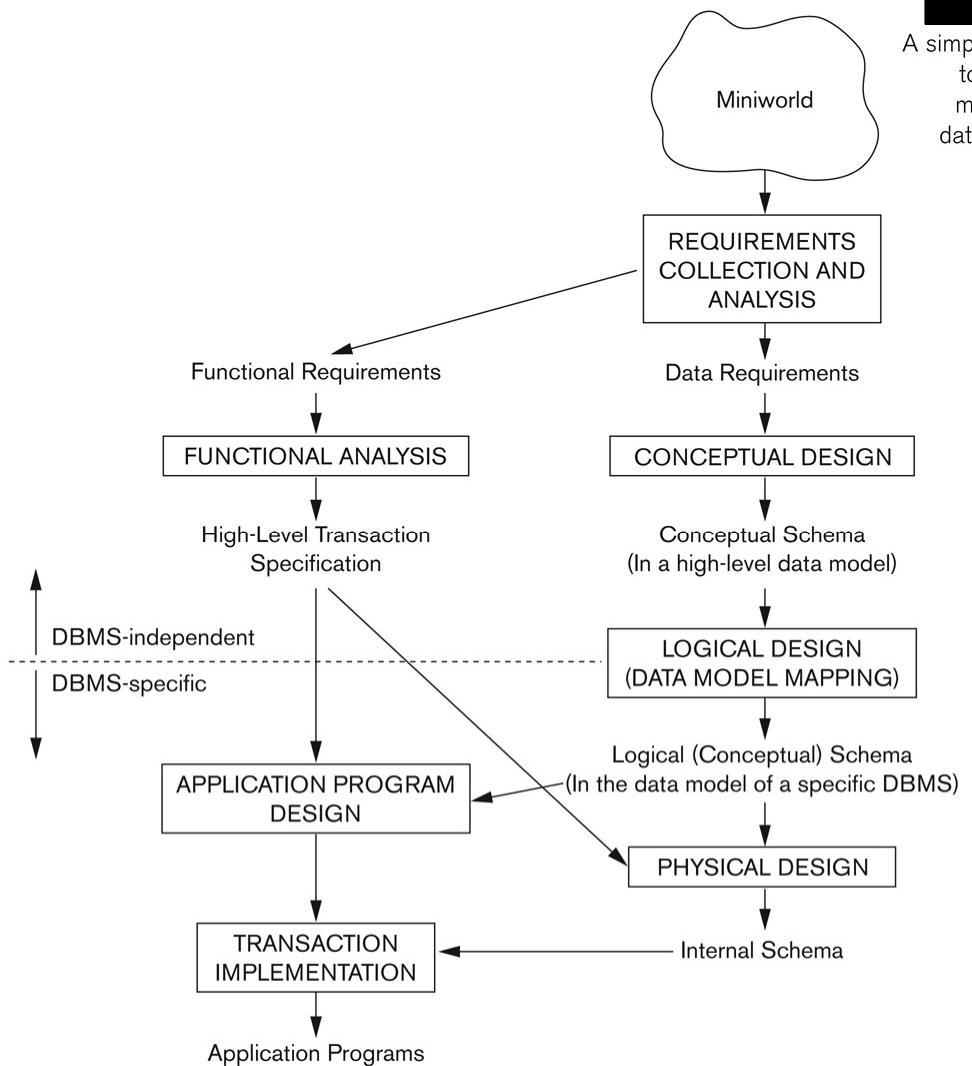
## **□ The New School for Data**

- Social Media, Mobile computing, Cloud computing and the internet produce Exabytes of primarily Unstructured Data on a daily basis
- Unstructured data has many potentially useful patterns (the case for Big Data Analytics)
- Structured Data still in use
- File Processing pushed back to front seat
- RDBMS still in use



# Overview of Database Design Process

A simplified diagram  
to illustrate the  
main phases of  
database design.

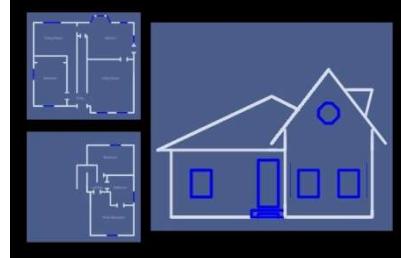


# How to build Database Application?

---

---

1. Blue print



2. Tools



3. Construction



4. House



1. Entity Relationship  
Diagram

2. SQL/UML/FD

3. Relations

4. Tables

# How to build Database Application?

---

## 1. Entity Relationship Diagram

- Entities
- Attributes
- Relationships



# How to build Database Application?

---

## 2. SQL/UML/FD

- SQL
  - DDL – Data Definition
  - DML – Data Manipulation
- UML
  - Notation for ER Diagram
- FD (Functional Dependency)
  - Update Anomalies
  - Delete Anomalies
  - Insert Anomalies



# How to build Database Application?

---

---

## 2. Relations

- Normalization
- Normal Forms
  - 1<sup>st</sup> Normal Form
  - 2<sup>nd</sup> Normal Form
  - 3<sup>rd</sup> Normal Form
  - BCNF



# How to build Database Application?

---

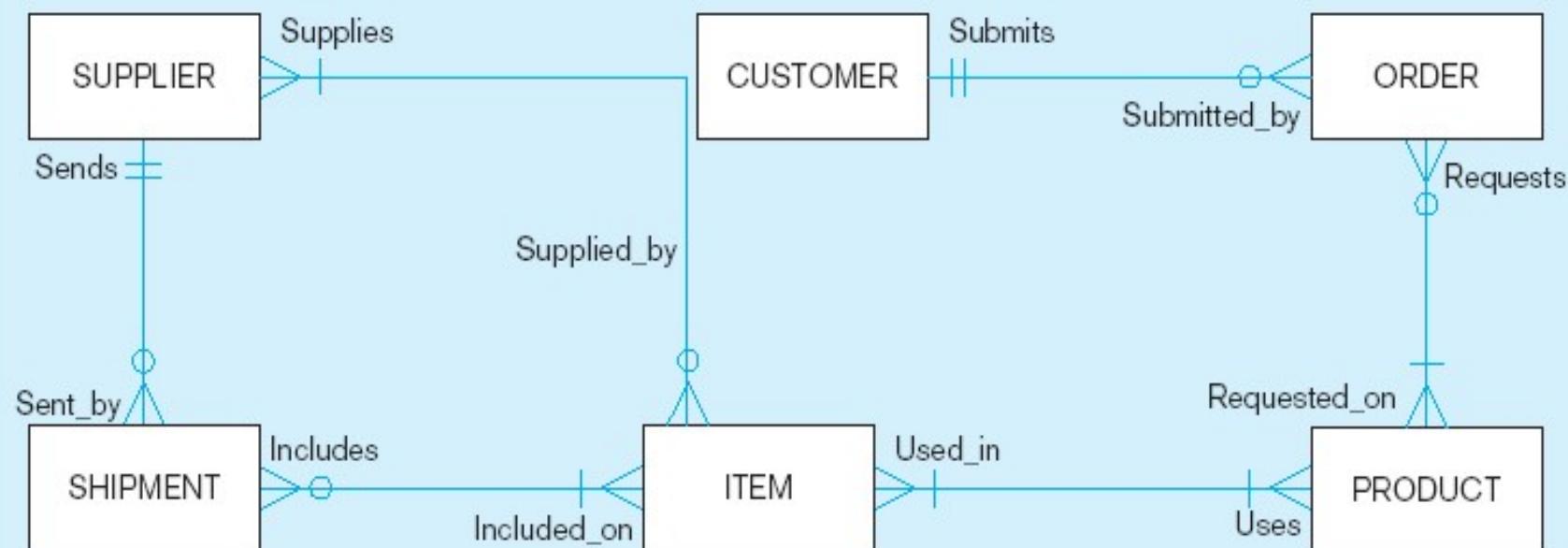
---

## 4. Tables

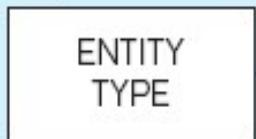
- Rows
- Columns
- Primary keys
- Foreign Keys
- Constraints



# Sample E-R Diagram



## Key



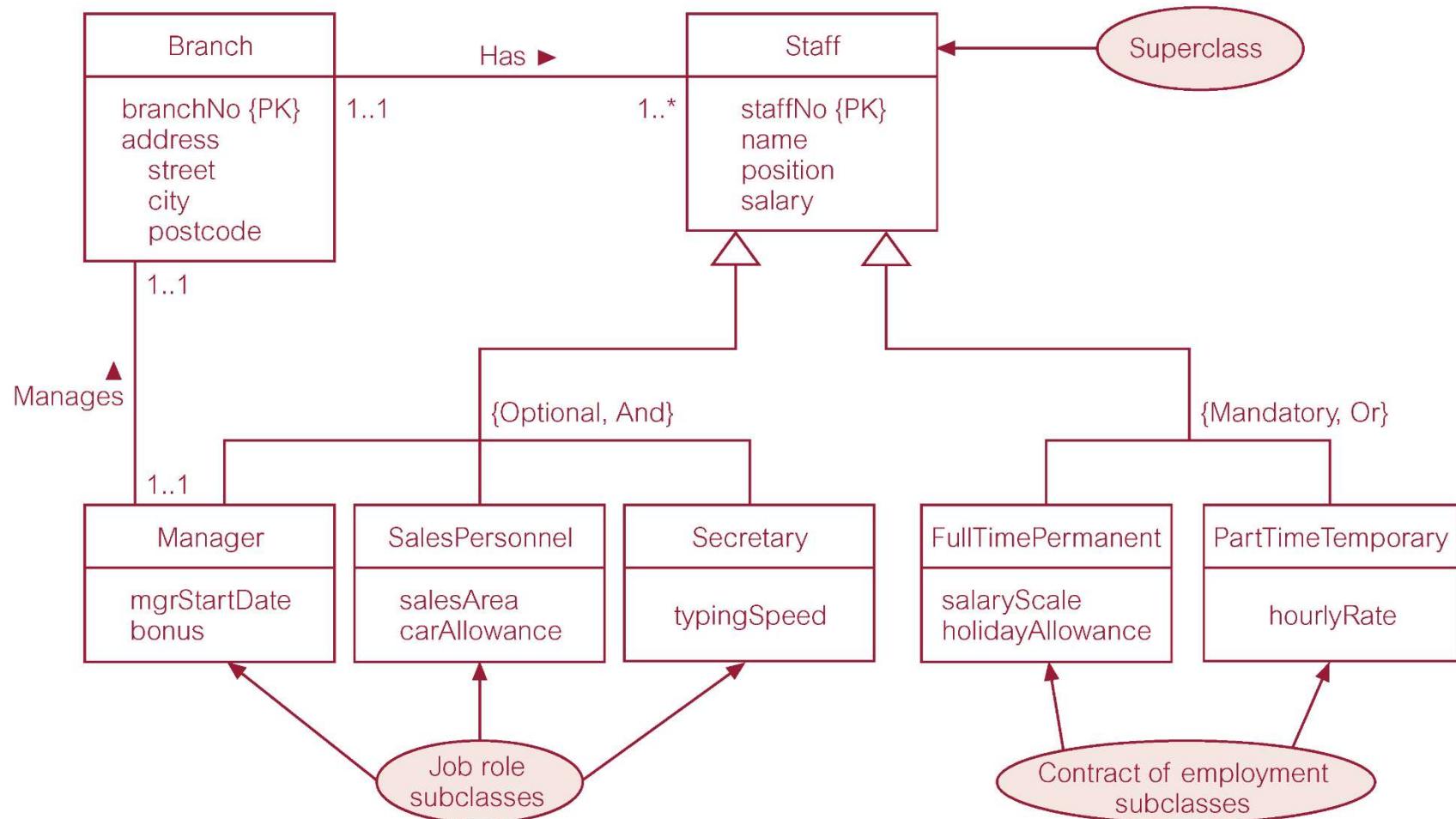
## Cardinalities



# AllStaff relation holding details of all staff

staffNo	name	position	salary	mgrStartDate	bonus	sales Area	car Allowance	typing Speed
SL21	John White	Manager	30000	01/02/95	2000			
SG37	Ann Beech	Assistant	12000					
SG66	Mary Martinez	Sales Manager	27000			SA1A	5000	
SA9	Mary Howe	Assistant	9000					
SL89	Stuart Stern	Secretary	8500					100
SL31	Robert Chin	Snr Sales Asst	17000			SA2B	3700	
SG5	Susan Brand	Manager	24000	01/06/91	2350			

# Specialization/generalization of Staff entity into job roles and contracts of employment



# Data , what is in the name?

---

---

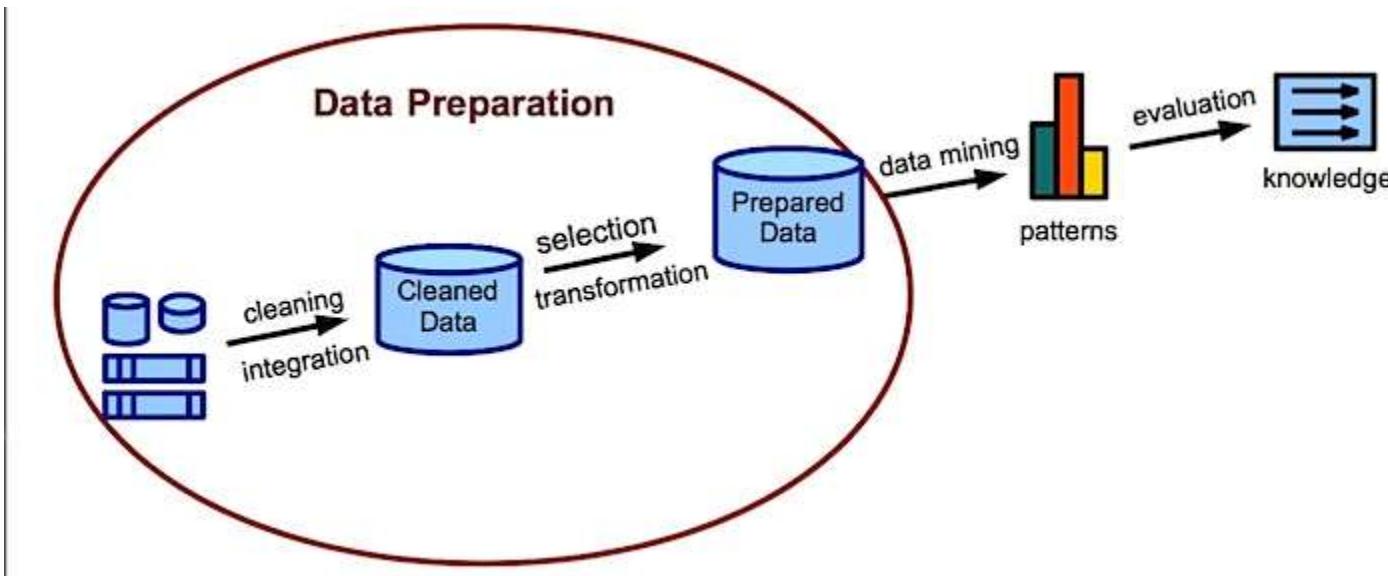
- Different terms used and interpreted differently:
  1. Data Preparation/Preprocessing
  2. Data Analysis
  3. Data Analytics
  4. Data Mining
  5. Data Processing
- Lets review each one of these terms ...



# Data , what is in the name?

## 1. Data Preparation/Preprocessing

- Data preparation (or data preprocessing) in this context means manipulation of data into a form suitable for further analysis and processing.



# Data , what is in the name?

---

---

## 2. Data Analysis

- Analysis proceeds design
- We say we do analysis to discover basic elements, relationships between the elements, and operations on the elements
- How we do analysis to design and build a database system for example?
- For example,
  - a company has employees, and offices
  - Company has name and budget
  - Office has number and address
  - Employee has an ID, name, salary
  - We want to be able to get a list of employees
  - We want to get a list of offices assigned to employees

# Data , what is in the name?

---

---

## 3. Data Analytics

- Is the science of examining raw data with the purpose of drawing conclusions about that information.
- Data Analytics use statistics, data mining, computer technology, etc to draw an inference
- Data analytics focuses on inference, the process of deriving a conclusion based solely on what is already known
- The term "analytics" has been used by many business intelligence (BI) software vendors as a buzzword to describe quite different functions
- Banks and credit cards companies, for instance, analyze withdrawal and spending patterns to prevent fraud or identity theft. Ecommerce companies examine Web site traffic or navigation patterns to determine which customers are more or less likely to buy a product or service based upon prior purchases or viewing trends

# Data , what is in the name?

---

---

## 4. Data Mining

- Is about sorting through large **data sets** using **software tools** and **Machine Learning algorithms** to identify **useful patterns** , **hidden knowledge**, and **hidden relationships**.

# Data , what is in the name?

---

---

## 5. Data Processing

- Apply Operations on data. (Addition, Multiplication, String tokenizer, etc.)

# **Big Data - Prime Time**



# “Big Data” is Growing

- 383+ Million Twitter accounts
  - 835+ Million Facebook subscribers
  - 1.2+ Billion Mobile Web users
  - Machine and sensor data
  - Over 6 million OnStar subscribers



This data as of 2012

# Big Data - Prime Time

---

---

- People spend over 500 billion minutes per month on Facebook.
- YouTube receives more than 2 billion viewers per day
- More than 30 billion pieces of content are shared each month on Facebook.
- Every minute, 24 hours of video is uploaded to YouTube
- As of December 2010, the average number of tweets sent per day was 110 million

# Big Data - Prime Time

---

---

- What Walmart reported ... BENTONVILLE, Ark., Nov. 23, 2012 – Today, Walmart U.S. reported its best ever Black Friday events.
  - The work of our associates is even more impressive when you consider they served approximately 22 million customers on Thursday.
  - Walmart's Black Friday plan included three events this year at 8 p.m., 10 p.m. and 5 a.m. During the high traffic period from 8 p.m. through midnight, Walmart processed nearly 10 million register transactions and almost 5,000 items per second.
  - Since its events began at 8 p.m., Walmart sold more than:
    - 1.8 million towels,
    - 1.3 million televisions,
    - 1.3 million dolls and
    - 250,000 bicycles.

# Big Data Prime Time

---

---

How much data you think roughly out there ?

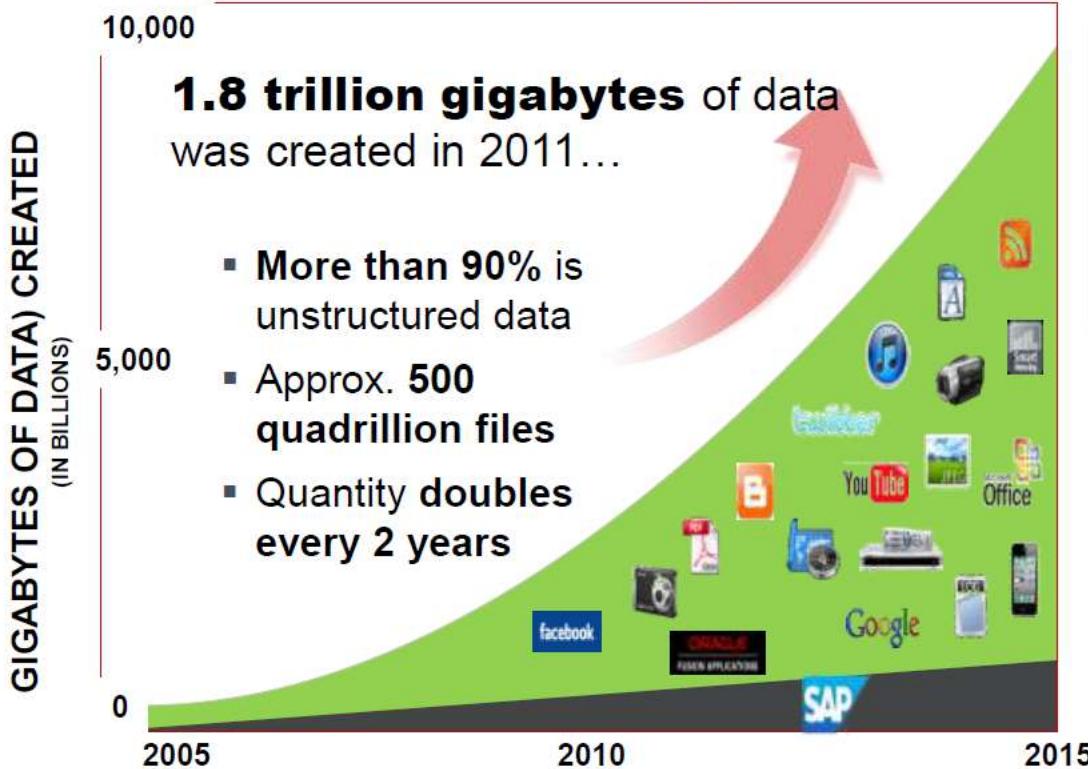
- How much data on your laptop?
- How much data on your PDA?
- How much data on your personal gmail account?
- How much data on facebook?
- How much data on twitter?
- etc. ...

Decimal	
Value	Metric
1000	kB kilobyte
$1000^2$	MB megabyte
$1000^3$	GB gigabyte
$1000^4$	TB terabyte
$1000^5$	PB petabyte
$1000^6$	EB exabyte
$1000^7$	ZB zettabyte
$1000^8$	YB yottabyte

1 EB = 1000000000000000000B =  $10^{18}$ bytes = 1000 petabytes = 1 billion gigabytes.

# Big Data Prime Time

“Big Data” → “Big Data Analytics”



*“There was 5 exabytes of information created between the dawn of civilization through 2003, but that much information is now created every 2 days, and the pace is increasing.”*

- Google CEO Eric Schmidt

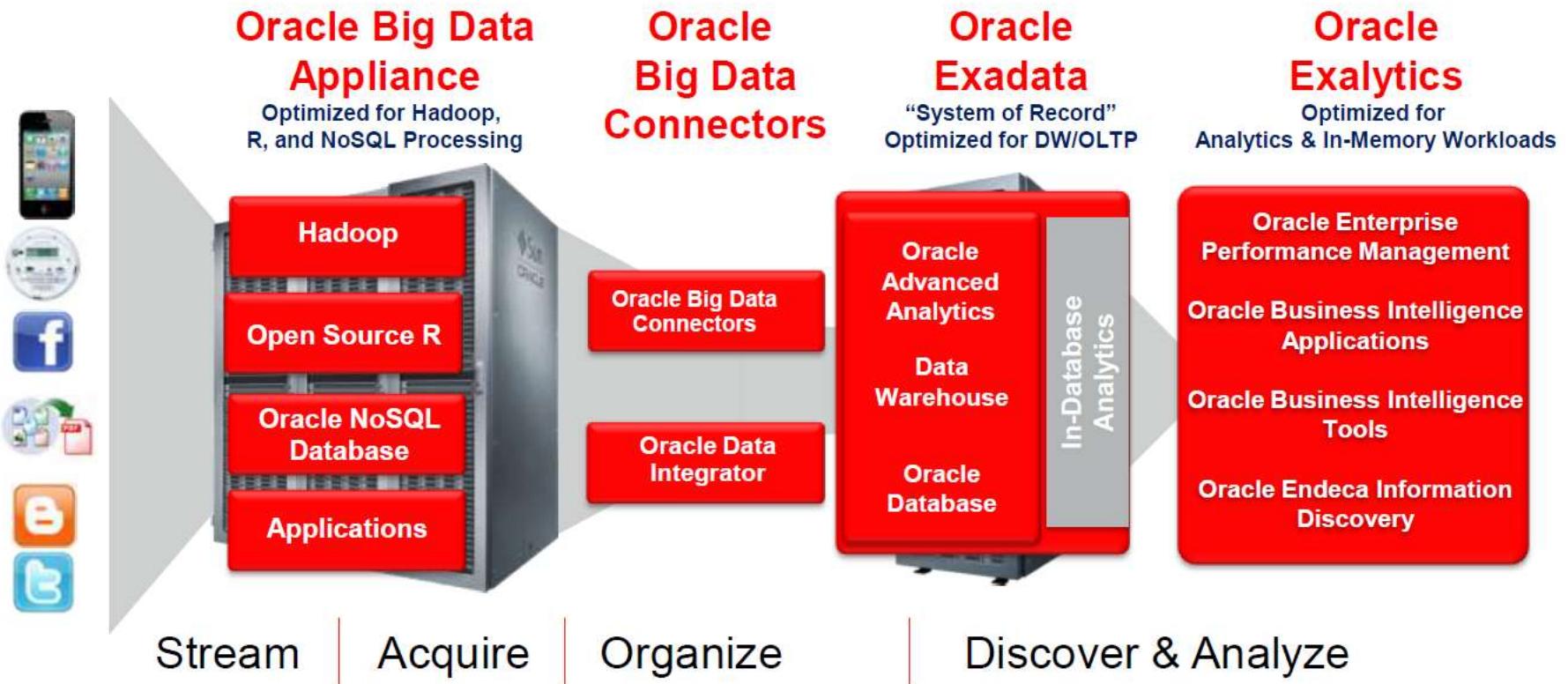
Requires capability to rapidly:

- ✓ Collect and integrate data
- ✓ Understand data & their relationships
- ✓ Respond and take action

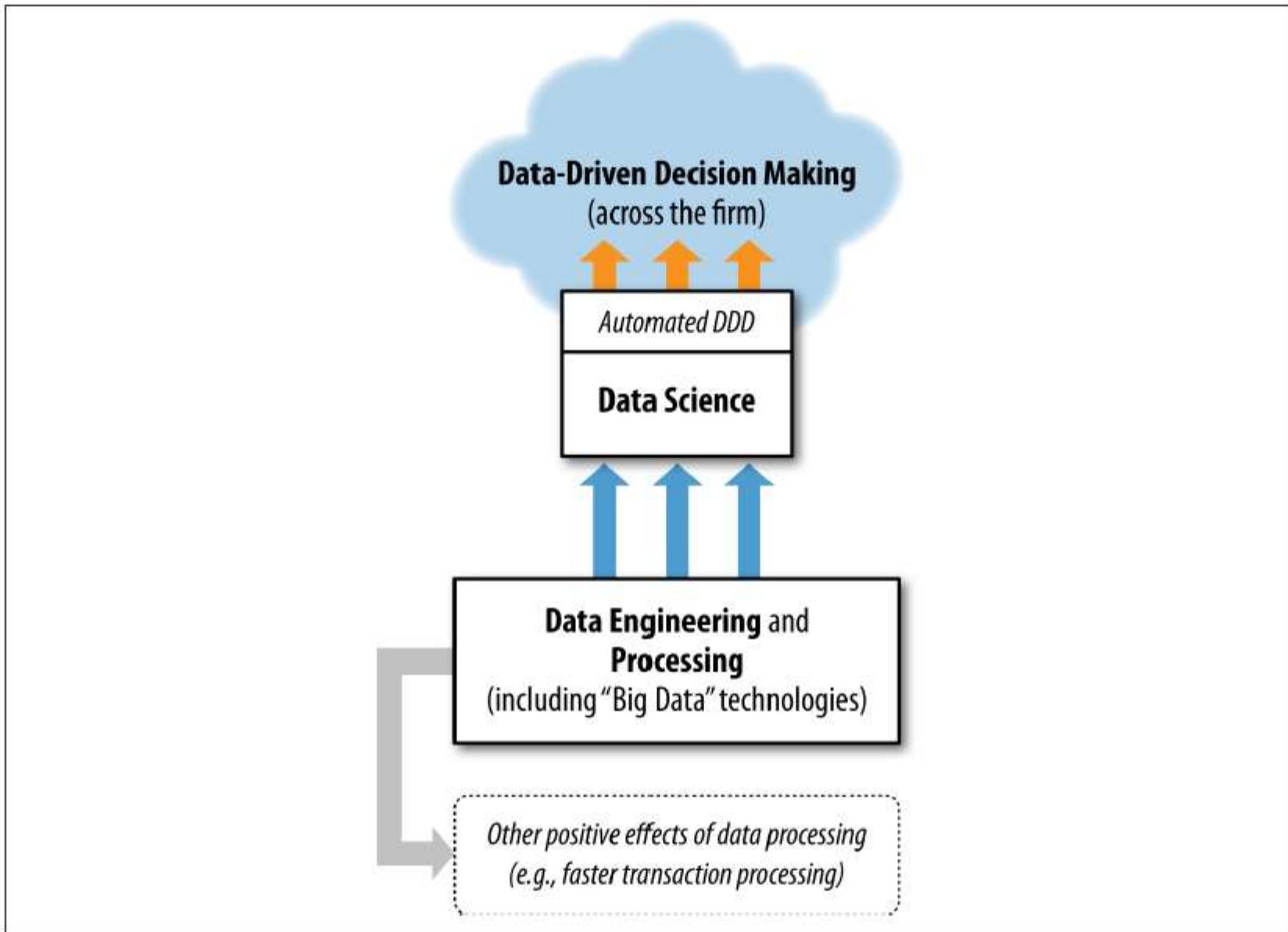
# Big Data Prime Time

Is RDBMS and Structured data history by now? No ...

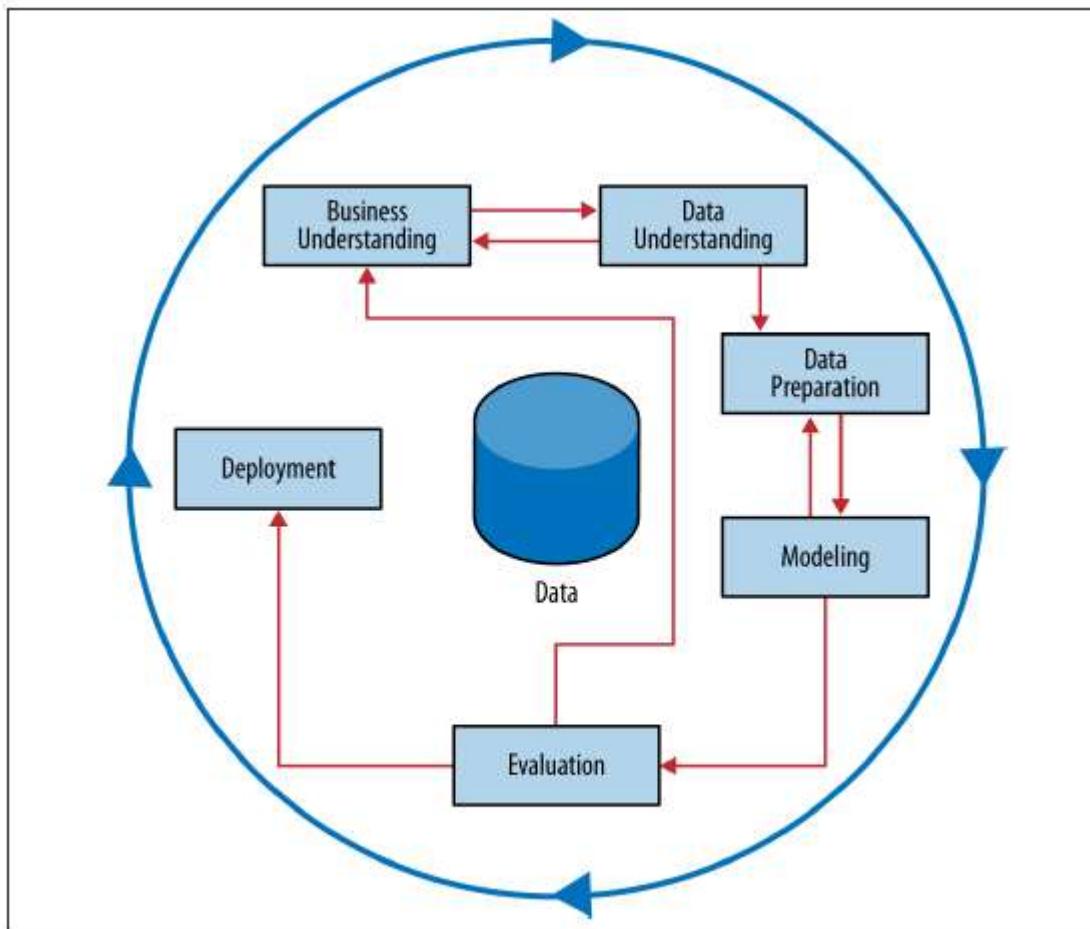
## Oracle Big Data Platform



# The Process

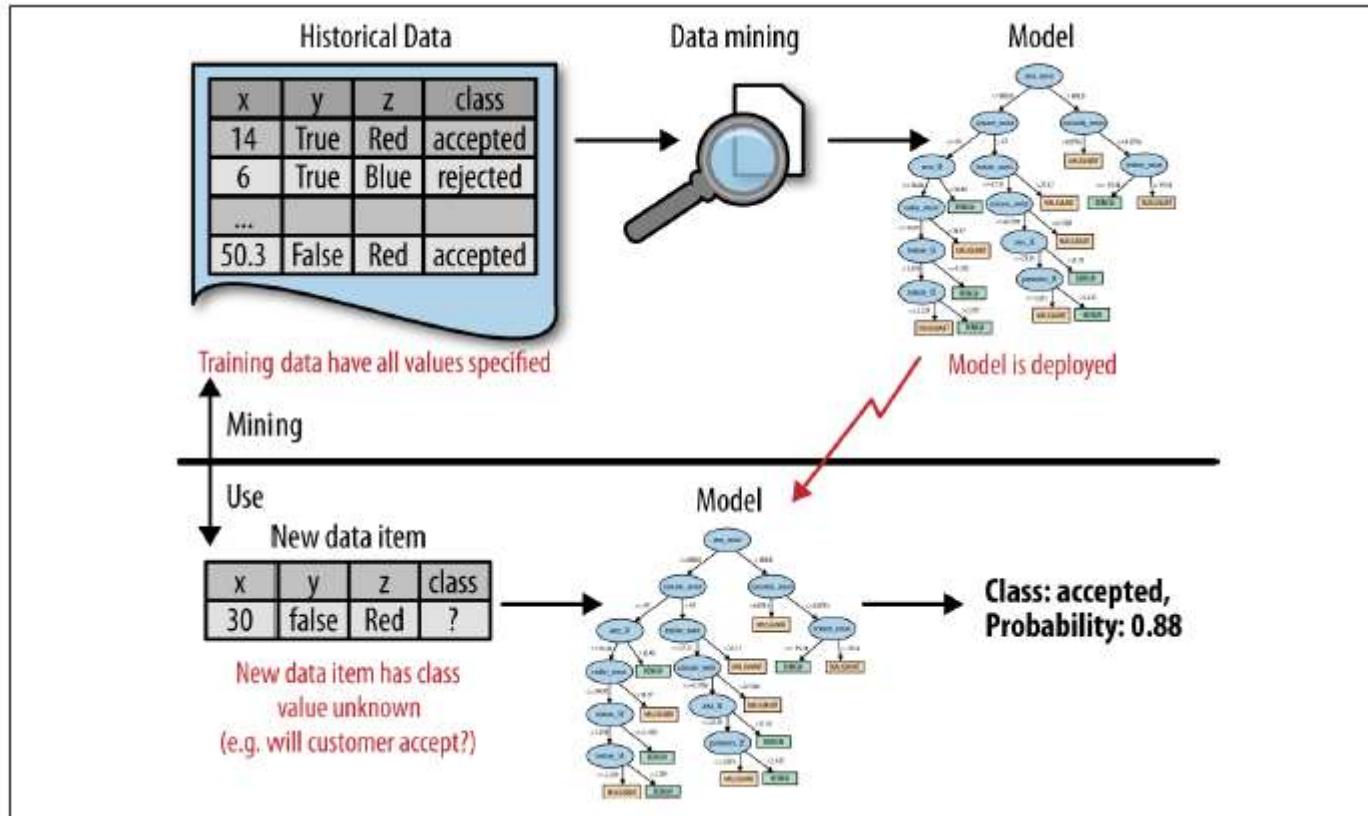


# The Process



*The CRISP data mining process.*

# The Process



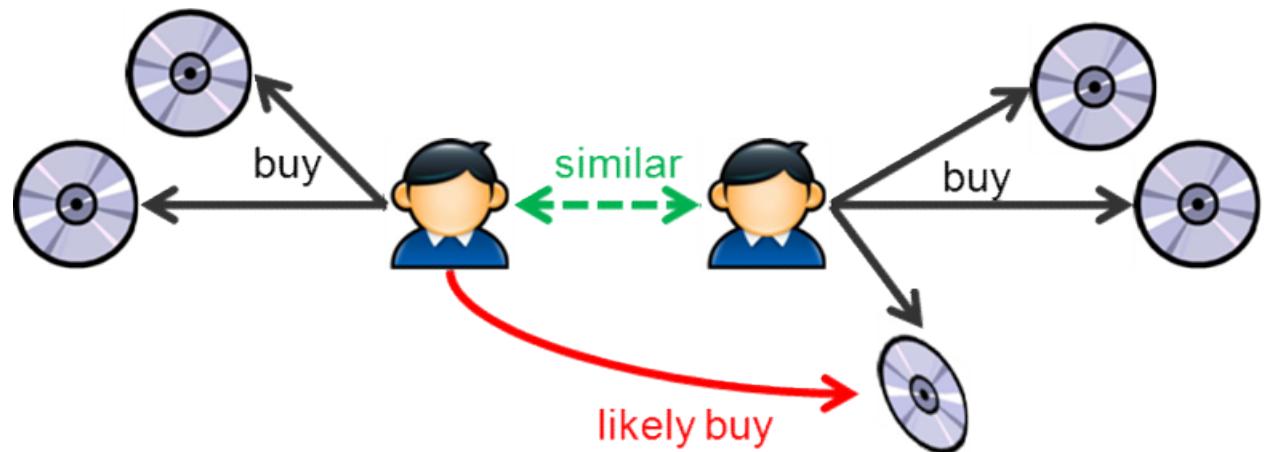
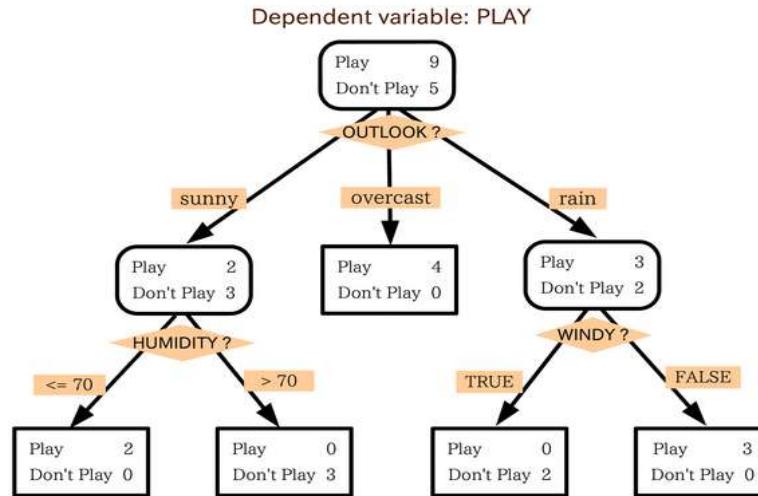
# Data Format

---

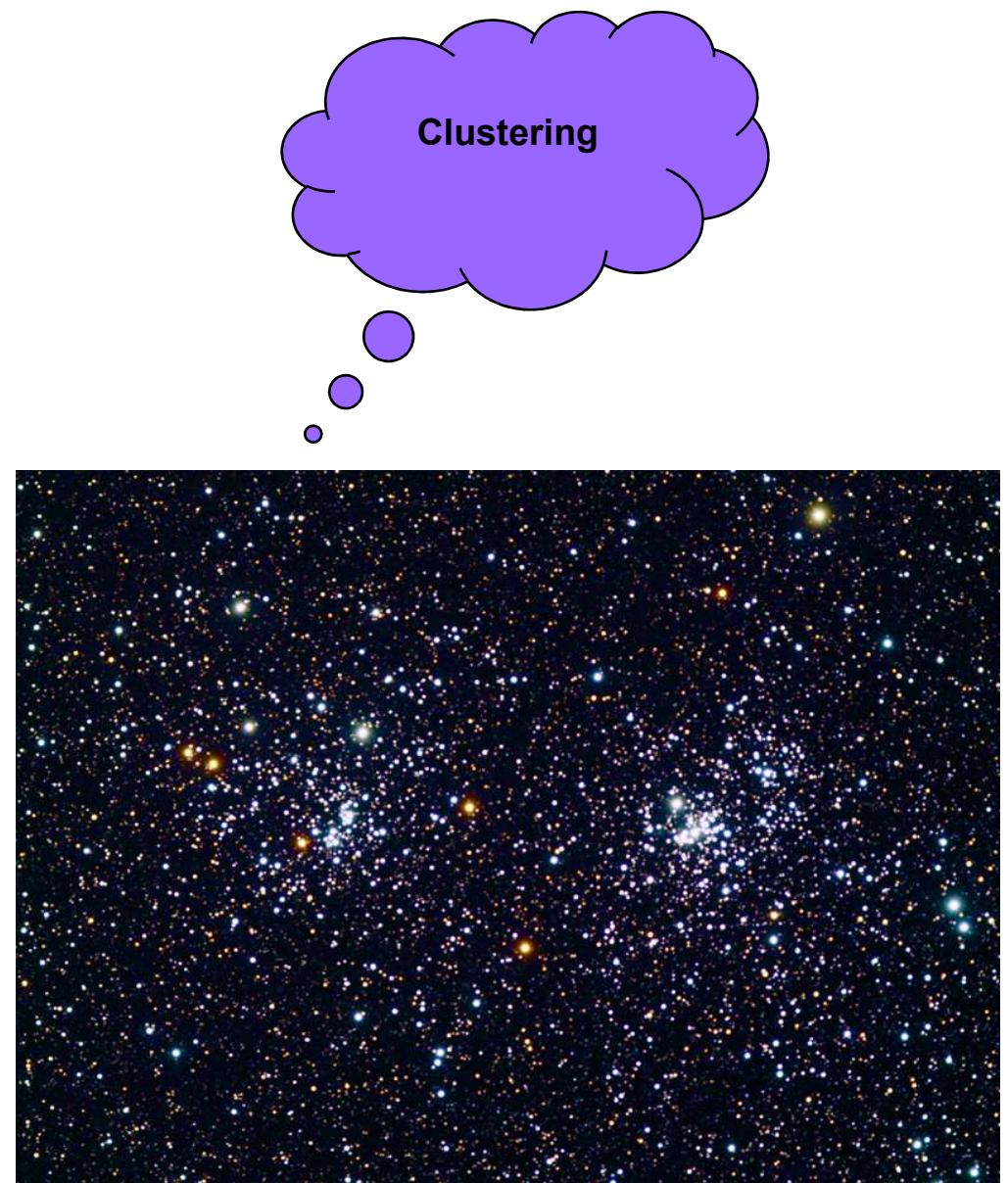
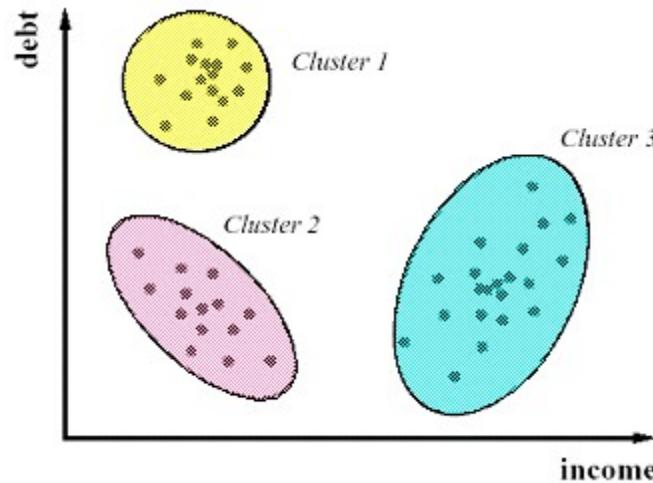
---

- Data is represented by ASCII code (American Standard Code for Information Interchange) which is the most widely used format.
- Data stored in Files as:
  1. Plaintext. Data is separated by comma, tab or space (plaintext). The most common extension is \*.csv (comma-separated value).
  2. Binary. Data is structured as a record by fixed blocks (formatted text)

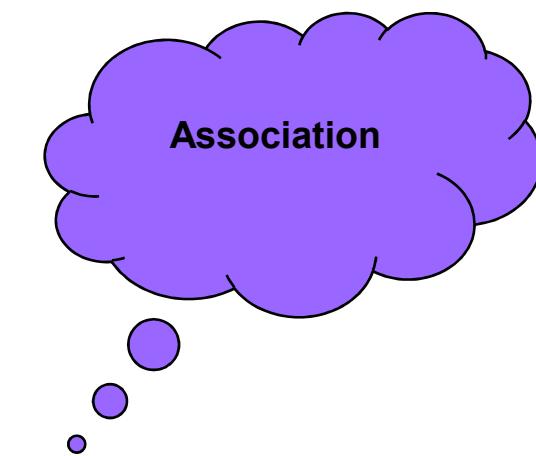
# Structured and Unstructured Data ...



# Structured and Unstructured Data ...



# Structured and Unstructured Data ...



**Take your TV  
to the next level**

[Shop all TV Accessories](#)

[TV Wall Mounts ▶](#)

[TV Stands & HDMI Cables ▶](#)

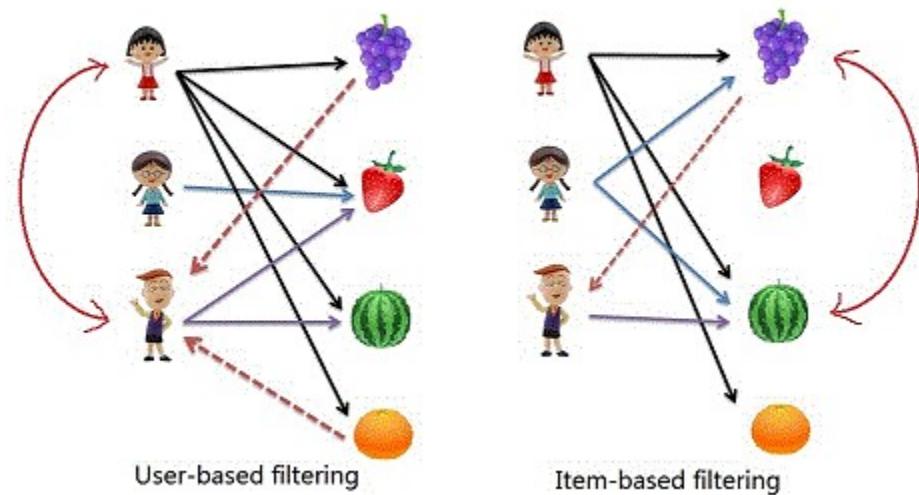
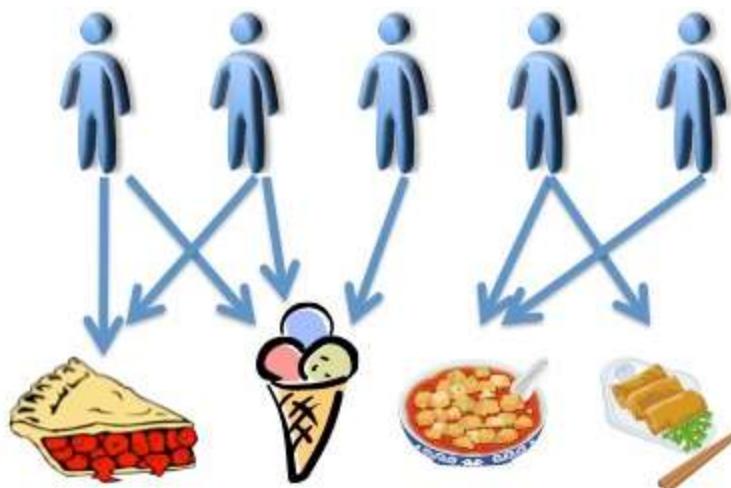
[Media Players ▶](#)

[Blu-ray Players ▶](#)

[Sound Bars ▶](#)

A promotional graphic for TV accessories. It features a central green rectangular area with the text "Take your TV to the next level" and a blue button below it labeled "Shop all TV Accessories". Surrounding this central area are five smaller rectangular boxes, each containing an image of a product and a link: "TV Wall Mounts", "TV Stands & HDMI Cables", "Media Players", "Blu-ray Players", and "Sound Bars".

# Structured and Unstructured Data ...



# Structured and Unstructured Data ...

More Items to Consider

You viewed Customers who viewed this also viewed

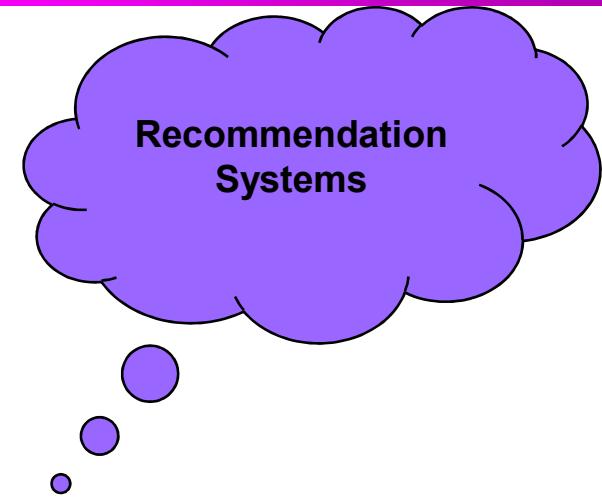
Similar Artists

Customers Who Viewed This Item Also Bought

Jupiter Years  
► Ted Paper! ★★★ \$24.95

Stanley Clarke & George Duke

The Diamond Sutra  
► Red Pine  
★★★★★ (20)  
Paperback  
\$13.57



PANDORA

New Station Type in artist, genre, or composer

Now Playing Music Feed My Profile

Lullabye  
Bio not available  
Sample of Artists on this Station

- The Piano Guys
- Jasmine Thompson
- Israel 'IZ' Kamakawiwo'ole
- Brian Crain

Play Station Like this genre station Share

People who also like this

# Technologies and Tools

---

- Ok, I got it, data come in different variants and we need to prepare and process data ...
- But what are the technologies and tools that we will use in this class



# python

The screenshot shows a web browser displaying the Python.org homepage. The page features a dark blue header with the Python logo and navigation links for Python, PSF, Docs, PyPI, Jobs, and Community. Below the header is a large banner with the Python logo and the word "python" in white. On the left, there's a code editor window showing Python code for arithmetic operations. To the right of the code is a section titled "Intuitive Interpretation" with text about Python's calculation rules. At the bottom of the banner, there's a call-to-action button with the text "Python is a programming language that lets you work quickly and integrate systems more effectively. [» Learn More](#)". The footer contains four main links: "Get Started", "Download", "Docs", and "Jobs".

File Edit View History Bookmarks Tools Help

Welcome to Python.org

Python Software Foundation (US) https://www.python.org

Search

Python PSF Docs PyPI Jobs Community

python™

# Python 3: Simple arithmetic

```
>>> 1 / 2
0.5
>>> 2 ** 3
8
>>> 17 / 3 # classic division returns a float
5.666666666666667
>>> 17 // 3 # floor division
5
```

Intuitive Interpretation

Calculations are simple with Python, and expression syntax is straightforward: the operators `+`, `-`, `*` and `/` work as expected; parentheses `()` can be used for grouping. [More about simple math functions in Python 3.](#)

1 2 3 4 5

Python is a programming language that lets you work quickly and integrate systems more effectively. [» Learn More](#)

Get Started

Whether you're new to

Download

Python source code and installers

Docs

Documentation for Python's

Jobs

Looking for work or have a Python

# Canopy

The screenshot shows a web browser window with the URL <https://store.enthought.com/downloads/#default>. The page is titled "Download Canopy" and features the Enthought logo. It provides download links for Windows, Linux, and Mac versions, both 64-bit and 32-bit, with a file size of 439.8 MB and an MD5 checksum. A brief description of Canopy as an integrated analysis environment for scientific computing, data analysis, and engineering is provided, along with a "FREE for all users" note about Canopy Express. A "DOWNLOAD Canopy" button is visible. To the right, there is a sidebar for "Python Training from the Pros" which includes a summary of Canopy's benefits, a link to "Online Courses", and a section for "Live Classroom Training".

File Edit View History Bookmarks Tools Help

Downloads - Enthought S... X +

https://store.enthought.com/downloads/#default

Search

PRODUCTS TRAINING CONSULTING COMPANY CONTACT

ENTHOUGHT SCIENTIFIC COMPUTING SOLUTIONS

Download Canopy

v. 1.7.4 · released July 21, 2016

Windows Linux Mac  
64-bit 32-bit

439.8 MB MD5: 06789a12d906fa83f6e1558615b722de

ENTHOUGHT CANOPY

Enthought Canopy: Easy Python Deployment Plus Integrated Analysis Environment for Scientific Computing, Data Analysis and Engineering

FREE for all users: Canopy Express, which includes access to 200+ of Canopy's most popular Python packages for scientific computing, data analysis, and engineering PLUS Canopy's integrated analysis environment. Get started today with easy deployment of pre-built, tested, and dependency-aware packages such as NumPy, SciPy, Pandas, Matplotlib, IPython and more.

Want to Get Even More From Canopy?

See our [subscription options](#) to unlock additional features such as:

- An extended library of over 450 pre-built and tested packages, with easy updates and customized package installation through Canopy's

DOWNLOAD Canopy

By downloading Canopy you acknowledge your acceptance of all the terms and conditions of the applicable license.

ENTHOUGHT TRAINING ON DEMAND See Courses

Python Training from the Pros

With Canopy you'll have a robust environment and tools for working in Python. Now learn how to maximize your results with training from Enthought's experts.

Online Courses

Online Python Training designed for the unique needs of scientists, engineers, analysts and quants.

Live Classroom Training

Open and custom classes at locations across the United

# SQL

File Edit View History Bookmarks Tools Help

W ISO 9075 - Wikipedia, the free encyclopedia +

en.wikipedia.org/wiki/ISO\_9075

Search

Create account Log in

Article Talk Read Edit View history Search

 WIKIPEDIA  
The Free Encyclopedia

Main page Contents Featured content Current events Random article Donate to Wikipedia Wikipedia store

Interaction Help About Wikipedia Community portal Recent changes Contact page

Tools What links here Related changes Upload file Special pages Permanent link Page information Wikidata item Cite this page

Print/export Create a book

## ISO 9075

From Wikipedia, the free encyclopedia

**ISO/IEC 9075 standard:** "Information technology - Database languages - SQL". See details by release:

- SQL:1999 is the *ISO/IEC 9075:1999* standard of 1999.
- SQL:2003 is the *ISO/IEC 9075:2003* standard of 2003.
- SQL:2006 is the *ISO/IEC 9075:2006* standard of 2006.
- SQL:2008 is the *ISO/IEC 9075:2008* standard of 2008.
- SQL:2011 is the *ISO/IEC 9075:2011* standard of 2011.

### External links [edit]

- "JTC 1/SC 32" technical committee

Categories: ISO standards

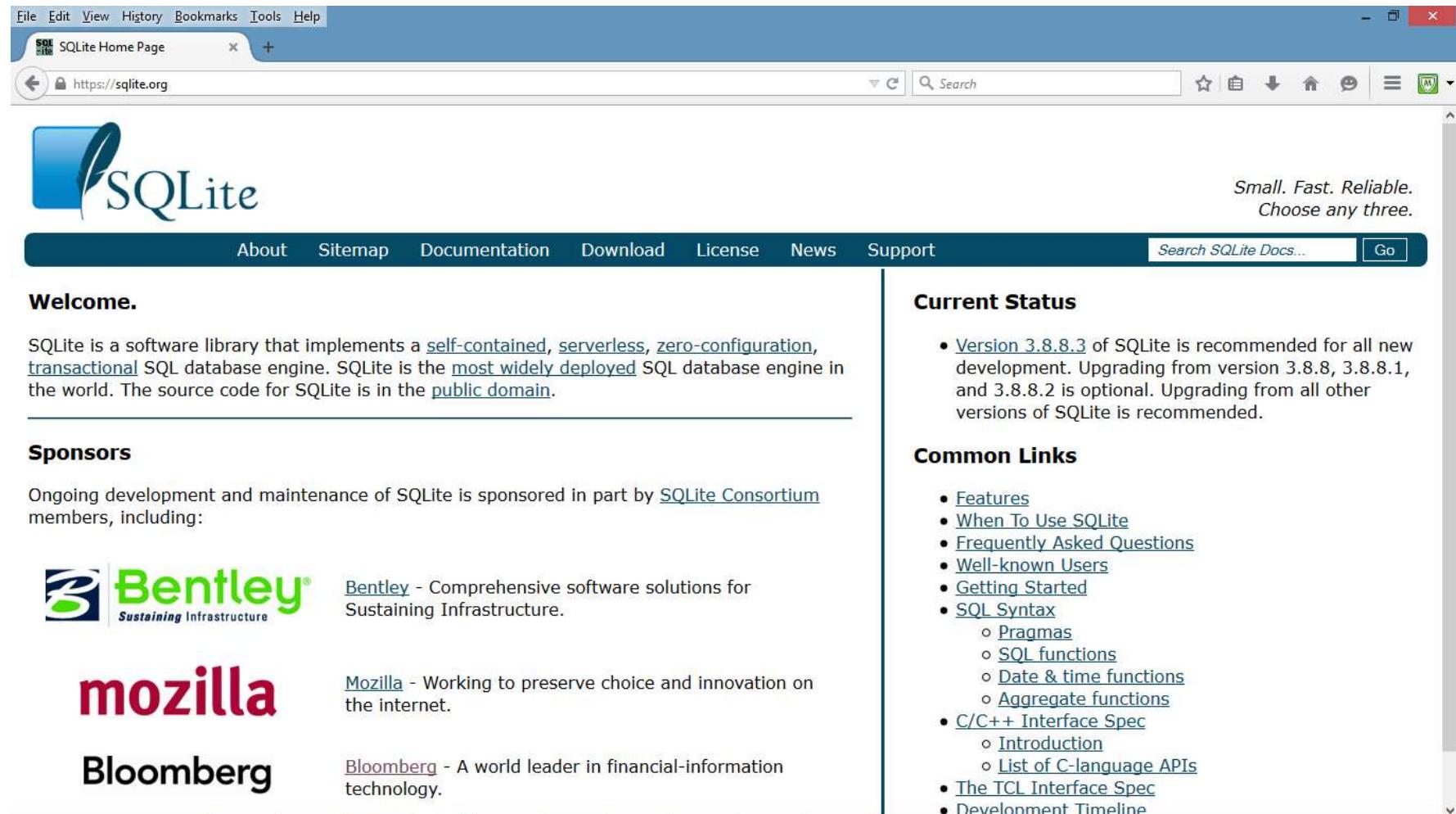
This page was last modified on 21 April 2013, at 04:49.

Text is available under the Creative Commons Attribution-ShareAlike License; additional terms may apply. By using this site, you agree to the Terms of Use and Privacy Policy. Wikipedia® is a registered trademark of the Wikimedia Foundation, Inc., a non-profit organization.

Privacy policy About Wikipedia Disclaimers Contact Wikipedia Developers Mobile view

# SQLite



The screenshot shows a web browser window displaying the SQLite Home Page. The page features a blue header bar with the SQLite logo and navigation links for About, Sitemap, Documentation, Download, License, News, and Support. A search bar at the top right allows users to search the SQLite documentation. The main content area includes a welcome message, information about the SQLite library, and sections for Sponsors and Current Status. The Sponsors section lists Bentley, Mozilla, and Bloomberg. The Current Status section highlights Version 3.8.8.3 as recommended for new development.

**Welcome.**

SQLite is a software library that implements a [self-contained](#), [serverless](#), [zero-configuration](#), [transactional](#) SQL database engine. SQLite is the [most widely deployed](#) SQL database engine in the world. The source code for SQLite is in the [public domain](#).

**Sponsors**

Ongoing development and maintenance of SQLite is sponsored in part by [SQLite Consortium](#) members, including:

**Bentley**  
Sustaining Infrastructure

**mozilla**

**Bloomberg**

**Current Status**

- [Version 3.8.8.3](#) of SQLite is recommended for all new development. Upgrading from version 3.8.8, 3.8.8.1, and 3.8.8.2 is optional. Upgrading from all other versions of SQLite is recommended.

**Common Links**

- [Features](#)
- [When To Use SQLite](#)
- [Frequently Asked Questions](#)
- [Well-known Users](#)
- [Getting Started](#)
- [SQL Syntax](#)
  - [Pragmas](#)
  - [SQL functions](#)
  - [Date & time functions](#)
  - [Aggregate functions](#)
- [C/C++ Interface Spec](#)
  - [Introduction](#)
  - [List of C-language APIs](#)
- [The TCL Interface Spec](#)
- [Development Timeline](#)

# PostgreSQL

The screenshot shows a web browser window displaying the PostgreSQL homepage. The URL in the address bar is [www.postgresql.org](http://www.postgresql.org). The page features a prominent blue header with the PostgreSQL logo and the text "The world's most advanced open source database". Below the header is a navigation menu with links for Home, About, Download, Documentation, Community, Developers, and Support. A main content area announces the release of PostgreSQL 9.4.4, 9.3.9, 9.2.13, 9.1.18 & 9.0.22. It includes a release date of June 12, 2015, and a note about bug fixes. There are links to "Release Announcement" and "Download". To the right, there are sections for "LATEST RELEASES" (with links to notes for each version), "SHORTCUTS" (links to Security, International Sites, Mailing Lists, Wiki, Report a Bug, and FAQs), and "SUPPORT US" (a call to support the project). A sidebar on the left highlights a "FEATURED USER" named Pascal Bouchareine from Gandi.net.

File Edit View History Bookmarks Tools Help

PostgreSQL: The world's m... +

www.postgresql.org

Search

Donate Contact Search Search

PostgreSQL

The world's most advanced open source database.

Home About Download Documentation Community Developers Support

**PostgreSQL 9.4.4, 9.3.9, 9.2.13, 9.1.18 & 9.0.22 Released!**

**12<sup>th</sup> June 2015**

The PostgreSQL Global Development Group is pleased to announce the availability of PostgreSQL 9.4.4, 9.3.9, 9.2.13, 9.1.18 & 9.0.22!

These new releases contain bug fixes over previous releases. All users should plan to upgrade their systems as soon as possible.

» [Release Announcement](#)  
» [Download](#)

» **FEATURED USER**

We've been using PostgreSQL for the Gandi IAAS/PAAS platform and recently internally, to build one of our live systems that stores/computes /outputs millions rows daily, very easily.

Pascal Bouchareine, [Gandi.net](#)

» **LATEST RELEASES**

9.4.4 · June 12, 2015 · [Notes](#)  
9.3.9 · June 12, 2015 · [Notes](#)  
9.2.13 · June 12, 2015 · [Notes](#)  
9.1.18 · June 12, 2015 · [Notes](#)  
9.0.22 · June 12, 2015 · [Notes](#)

[Download](#) | [RSS](#)  
[Why should I upgrade?](#)

» **SHORTCUTS**

» [Security](#)  
» [International Sites](#)  
» [Mailing Lists](#)  
» [Wiki](#)  
» [Report a Bug](#)  
» [FAQs](#)

» **SUPPORT US**

PostgreSQL is free. Please support our work by making a [donation](#).

# Bottom Line ...

---

---

## 1. What you will learn in this class?

- Learn the fundamental concepts for data management, data preparation, relational database, and file processing
- Database Engines
- Programming: Python and SQL

## 2. What you will do in this class?

- Assignments (covering different topics)
- Term project
- Exercises (Python programming - data preparation/processing)

# Bottom Line ...

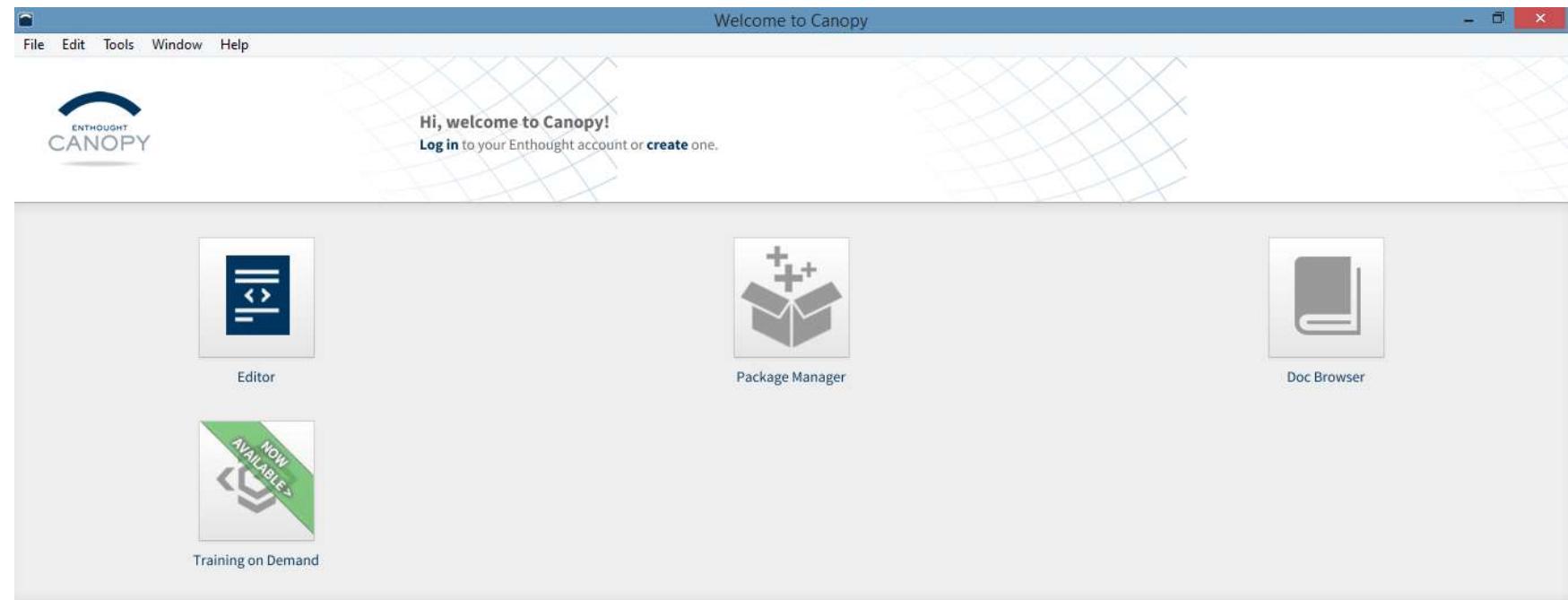
---

---

## 3. Logistics?

- Participate on a weekly basis in EVERY thread posted by me in the class discussions on Canvas; deadline for every week is 11:59pm Sunday night
- I will be posting RECORDED sync sessions almost on a weekly basis.
- Your attendance is NOT REQUIRED for any sync session that I might hold; (NO points-credit for attending sync sessions)
- Whenever there is a live sync session, I will be sending an email in advance
- ALL SYNC SESSIONS ARE RECORDED so if you can't attend any LIVE sync session you can go to Canvas and watch theses recordings on your free time
- Please do NOT discuss the class assignments on Canvas. If you have a doubt about an assignment please EMAIL me

# Canopy



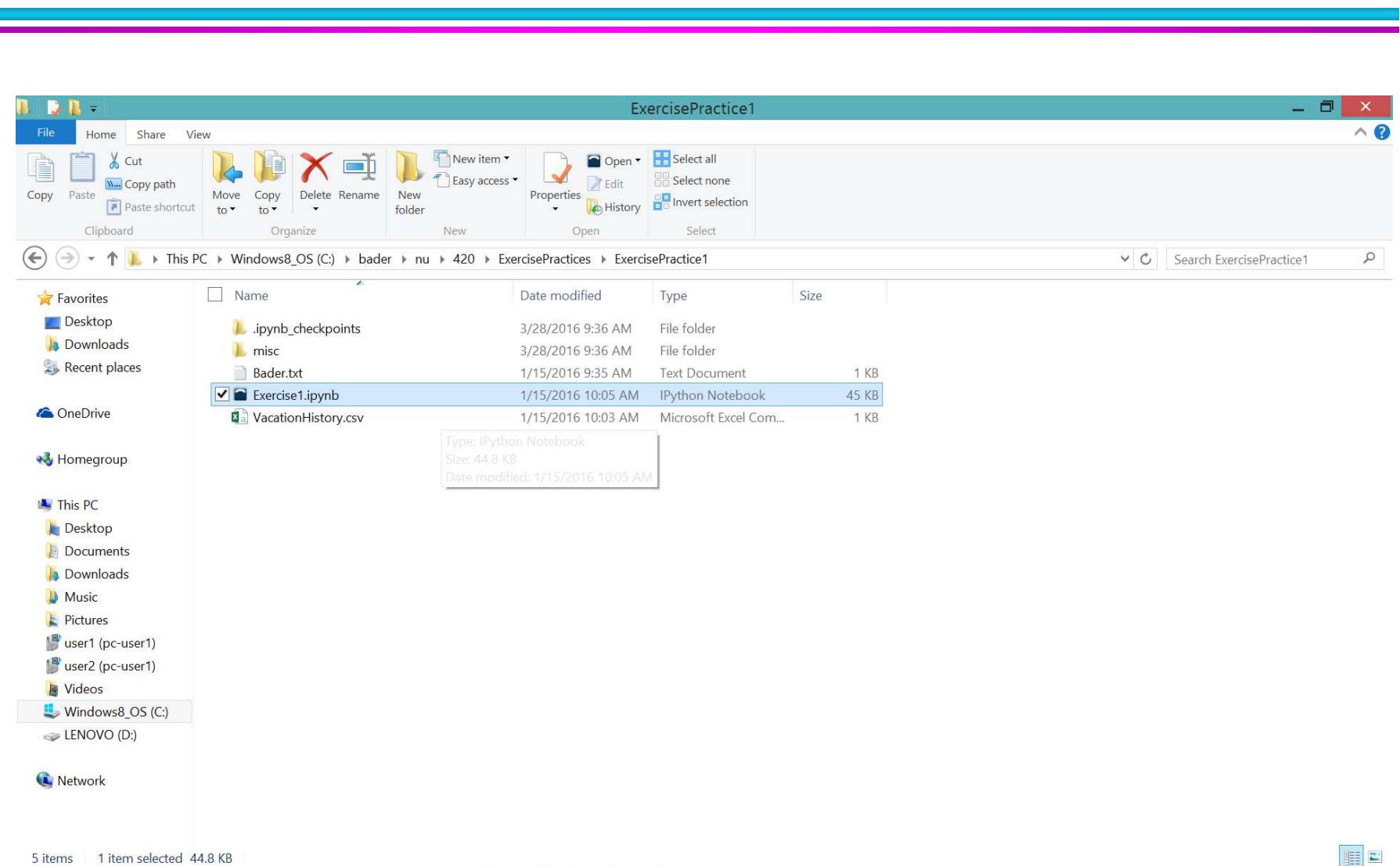
## Recent files

Exercise1.ipynb  
Practice1.py  
Exercise1.ipynb  
Practice1.py  
ep1.ipynb

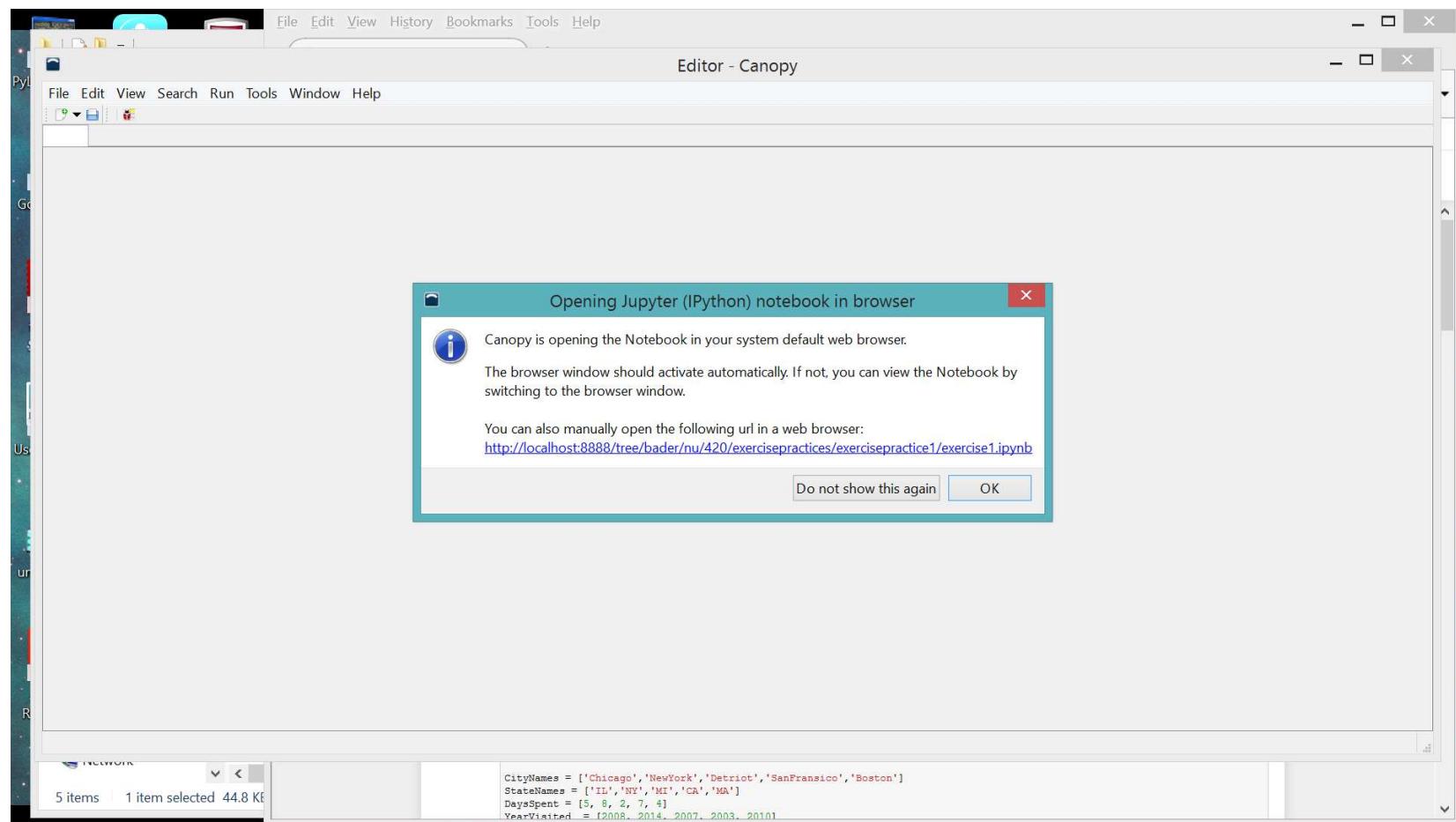
Open an existing file  Restore previous session

Version: 1.5.2.2785  
No updates found.

# Ipython Notebook



# Ipython Notebook



# Ipython Notebook

The screenshot shows a Jupyter Notebook interface with the following details:

- Header:** File, Edit, View, History, Bookmarks, Tools, Help.
- Title Bar:** exercise1, localhost:8888/notebooks/bader/nu/420/exercisepractices/exercisepractice1/exercise1.ipynb, Search, and various browser icons.
- Toolbar:** File, Edit, View, Insert, Cell, Kernel, Help, and a Cell Toolbar dropdown set to None.
- Kernel:** Python 2.
- Section Header:** Exercise Practice #1.
- Text Content:** This exercise will present a simple walkthrough of the Python script structure and what could be done when using Python for the purposes of data analysis and data preparation. Python is an interpreted programming language that has many packages designed primarily for data analysis and preparation.  
Data munging/wrangling is made simpler when using Python compared to R. And Python has many packages that can help you do the task, following is the list of the major Python packages that will be used in this class:  
1. Pandas  
2. numpy  
3. cPickle  
4. SQLAlchemy  
5. PyMongo
- Code Cell 1:** In [1]:

```
# General syntax to import specific functions in a library:  
##from (library) import (specific library function)  
  
from pandas import DataFrame, read_csv
```
- Code Cell 2:** In [2]:

```
#General syntax to import a library but no functions:  
##import (library) as (give the library a nickname/alias)  
  
import matplotlib.pyplot as plt  
import pandas as pd  
import sys # needed to determine Python version number
```
- Code Cell 3:** In [3]:

```
print 'Python version ' + sys.version  
print 'Pandas version ' + pd.__version__
```

Output:  
Python version 2.7.6 | 64-bit | (default, Sep 15 2014, 17:36:35) [MSC v.1500 64 bit (AMD64)]  
Pandas version 0.16.2

# Ipython Notebook

The screenshot shows a Jupyter Notebook interface with the following details:

- Header:** File Edit View History Bookmarks Tools Help
- Title Bar:** exercise1
- Toolbar:** Includes icons for file operations, search, and navigation.
- Section Header:** Jupyter exercise1 Last Checkpoint: 04/02/2015 (autosaved)
- Cell Toolbar:** Set to "None".
- Cell Type:** Python 2
- Content Area:** Contains three code cells labeled In [1], In [2], and In [3].
  - In [1]:**

```
# General syntax to import specific functions in a library:  
##from (library) import (specific library function)  
  
from pandas import DataFrame, read_csv
```
  - In [2]:**

```
#General syntax to import a library but no functions:  
##import (library) as (give the library a nickname/alias)  
  
import matplotlib.pyplot as plt  
import pandas as pd  
import sys # needed to determine Python version number
```
  - In [3]:**

```
print 'Python version ' + sys.version  
print 'Pandas version ' + pd.__version__
```

Output: Python version 2.7.6 | 64-bit | (default, Sep 15 2014, 17:36:35) [MSC v.1500 64 bit (AMD64)]
- Bottom Status Bar:** localhost:8888/notebooks/bader/nu/420/exercisepractices/exercisepractice1/exercise1.ipynb#

# Ipython Notebook

The screenshot shows a Jupyter Notebook interface with the following details:

- Header:** File Edit View History Bookmarks Tools Help
- Title Bar:** exercise1
- Toolbar:** Back, Forward, Refresh, Search, Python 2
- Section Header:** Jupyter exercise1 Last Checkpoint: 04/02/2015 (autosaved)
- Cell 4:** In [4]:

```
#####
# 1. Create Data #####
#####
CityNames = ['Chicago','NewYork','Detroit','SanFransico','Boston']
StateNames = ['IL','NY','MI','CA','MA']
DaysSpent = [5, 8, 2, 7, 4]
YearVisited = [2008, 2014, 2007, 2003, 2010]

VacationDataSet = zip(CityNames ,StateNames, DaysSpent, YearVisited)

df = pd.DataFrame(data = VacationDataSet, columns=['CityNames', 'StateNames', 'DaysSpent', 'YearVisited'])

#Sanity test: what is in the dataframe object
df
```
- Cell 4 Output:** Out[4]:

	CityNames	StateNames	DaysSpent	YearVisited
0	Chicago	IL	5	2008
1	NewYork	NY	8	2014
2	Detroit	MI	2	2007
3	SanFransico	CA	7	2003
4	Boston	MA	4	2010
- Cell 5:** In [5]:

```
#####
# 2. Write Data to CSV File #####
#####
df.to_csv('VacationHistory.csv',index=False,header=False)
```
- Cell 6:** In [6]:

```
#####
# 3. Read Data From CSV File #####
#####
```

# Ipython Notebook

The screenshot shows an Ipython Notebook interface with a teal header bar and a white main area. The header bar includes a menu bar with File, Edit, View, History, Bookmarks, Tools, and Help. Below the menu is a tab bar with 'exercise1' selected. The main area displays two code cells and their resulting plots.

**Code Cell 1:**

```
%matplotlib inline

df['DaysSpent'].plot()

ax = plt.gca()
ax.grid(True)
ax.get_yaxis().get_major_formatter().set_useOffset(False)
```

**Plot 1:** A line graph showing Days Spent over time. The x-axis ranges from 0.0 to 4.0, and the y-axis ranges from 2 to 8. The plot shows a triangular wave pattern with peaks at approximately (0.5, 5), (1.0, 8), (3.0, 7), and (3.5, 4).

**Code Cell 2:**

```
In [15]: df['YearVisited'].plot()

ax = plt.gca()
ax.grid(True)
ax.get_yaxis().get_major_formatter().set_useOffset(False)
```

**Plot 2:** A line graph showing Year Visited over time. The x-axis ranges from 2004 to 2014, and the y-axis ranges from 2004 to 2014. The plot shows a triangular wave pattern with peaks at approximately (2008, 2008), (2010, 2012), (2012, 2014), and (2014, 2008).

## What you need to submit for Exercise #1 on Canvas?

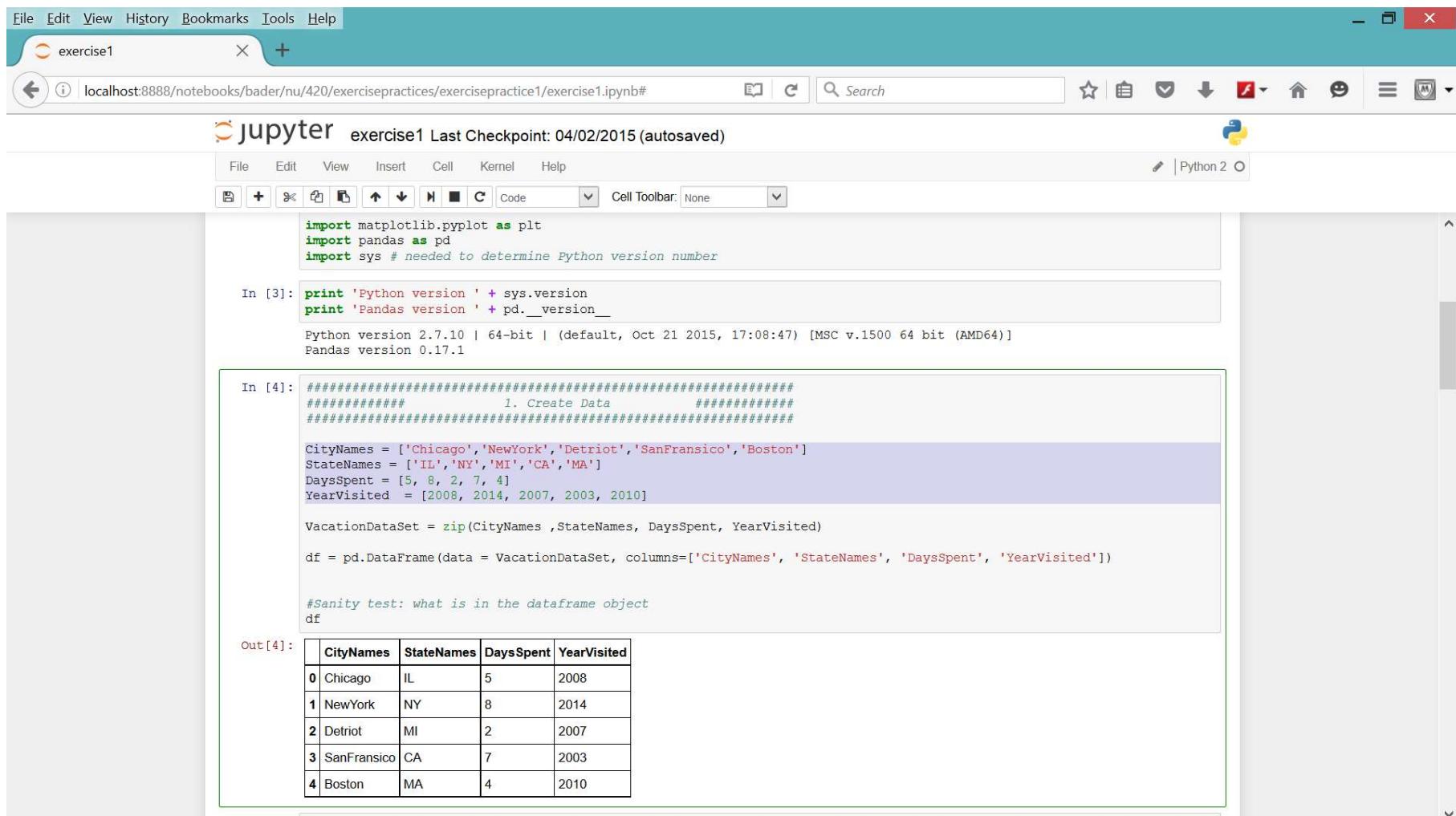
---

---

1. Change the data in Exercise #1 to reflect the cities/states/length of your past 5 vacations
2. Run the IPython Notebook Script Again
3. Save your updated IPython Notebook Script
4. Submit your updated IPython Notebook Script on Canvas

# What you need to submit for Exercise #1 on Canvas?

- The following is the only code you need to update



The screenshot shows a Jupyter Notebook interface with the following details:

- Header:** File Edit View History Bookmarks Tools Help
- Title Bar:** exercise1
- Toolbar:** Includes icons for back, forward, search, and various notebook operations.
- Header Bar:** Jupyter exercise1 Last Checkpoint: 04/02/2015 (autosaved)
- File Menu:** File Edit View Insert Cell Kernel Help
- Cell Toolbar:** Code Cell Toolbar: None
- In [3]:** Python version 2.7.10 | 64-bit | (default, Oct 21 2015, 17:08:47) [MSC v.1500 64 bit (AMD64)]  
Pandas version 0.17.1
- In [4]:** Code block containing:

```
#####
##### 1. Create Data #####
#####

CityNames = ['Chicago','NewYork','Detroit','SanFransico','Boston']
StateNames = ['IL','NY','MI','CA','MA']
DaysSpent = [5, 8, 2, 7, 4]
YearVisited = [2008, 2014, 2007, 2003, 2010]

VacationDataSet = zip(CityNames ,StateNames, DaysSpent, YearVisited)

df = pd.DataFrame(data = VacationDataSet, columns=['CityNames', 'StateNames', 'DaysSpent', 'YearVisited'])

#Sanity test: what is in the dataframe object
df
```
- Out[4]:** Dataframe output showing the following data:

	CityNames	StateNames	DaysSpent	YearVisited
0	Chicago	IL	5	2008
1	NewYork	NY	8	2014
2	Detroit	MI	2	2007
3	SanFransico	CA	7	2003
4	Boston	MA	4	2010