MSDS420

Atef Bader, PhD

Agenda

- Syllabus
- Canvas – Course Homepage
- FAQ
- Course Topics - Walkthrough
- Technologies and Tools: Python, SQlite, Anaconda, PostreSQL (SQL) and ElasticSearch (NoSQL)

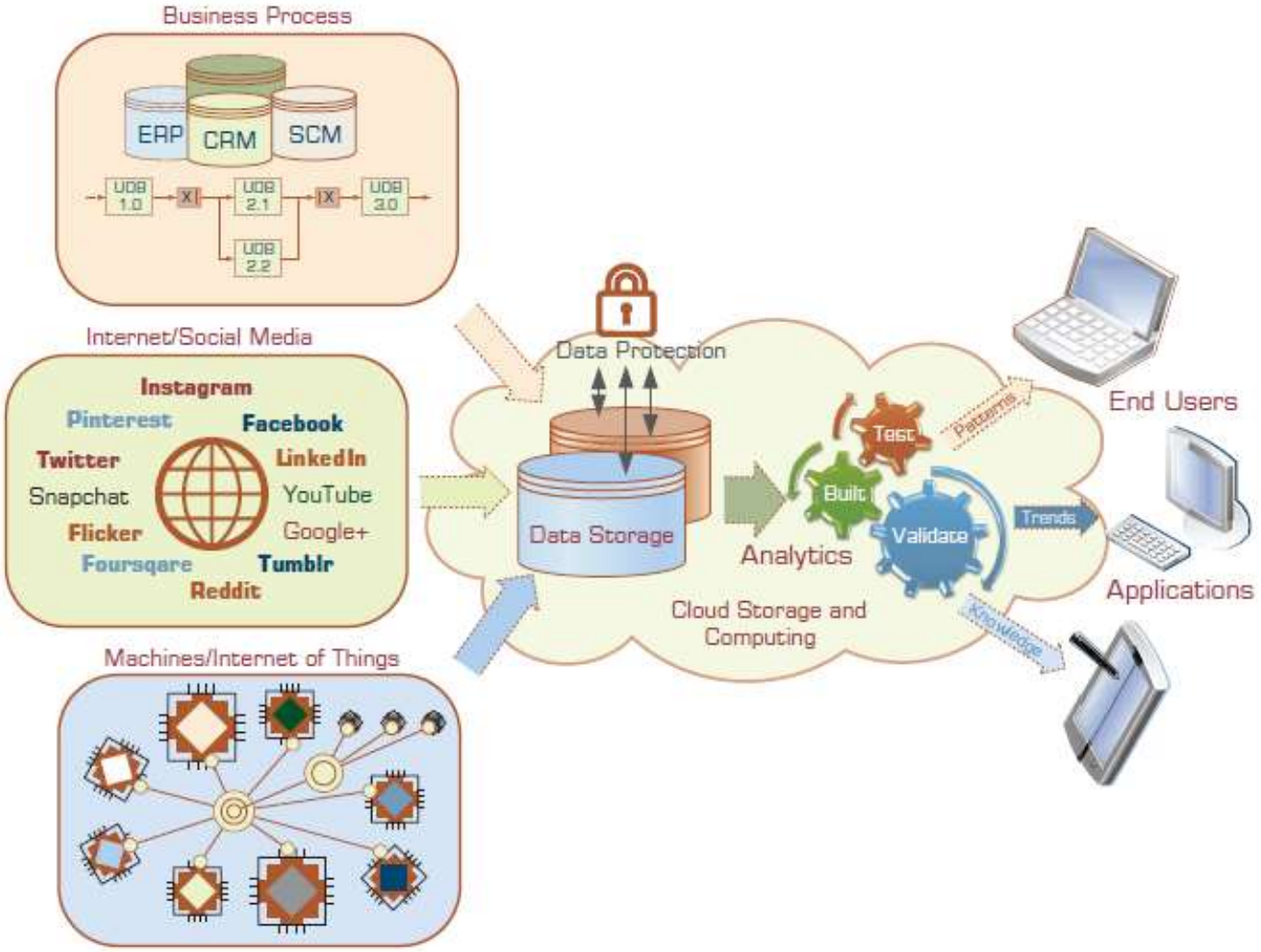
FAQ

- 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
  - Database Engines: SQLite, PostgreSQL, ElasticSearch
- Do I need programming background?
  - Yes, if you programmed in Python, R, Java, JavaScript, C, or C# before, you can manage in this class
  - If you didn't program before in any high-level programming language, drop the class, and take MSDS430 (Python for Data Science) first.
- Where can I find the due dates for assignments?
  - Canvas/Syllabus

The Nature of Data

- Data: a collection of facts
  - usually obtained as the result of experiences, observations, or experiments
- Data may consist of numbers, words, images, …
- Data is the lowest level of abstraction (from which information and knowledge are derived)
- Data is the source for information and knowledge
- Data quality and data integrity → critical to analytics

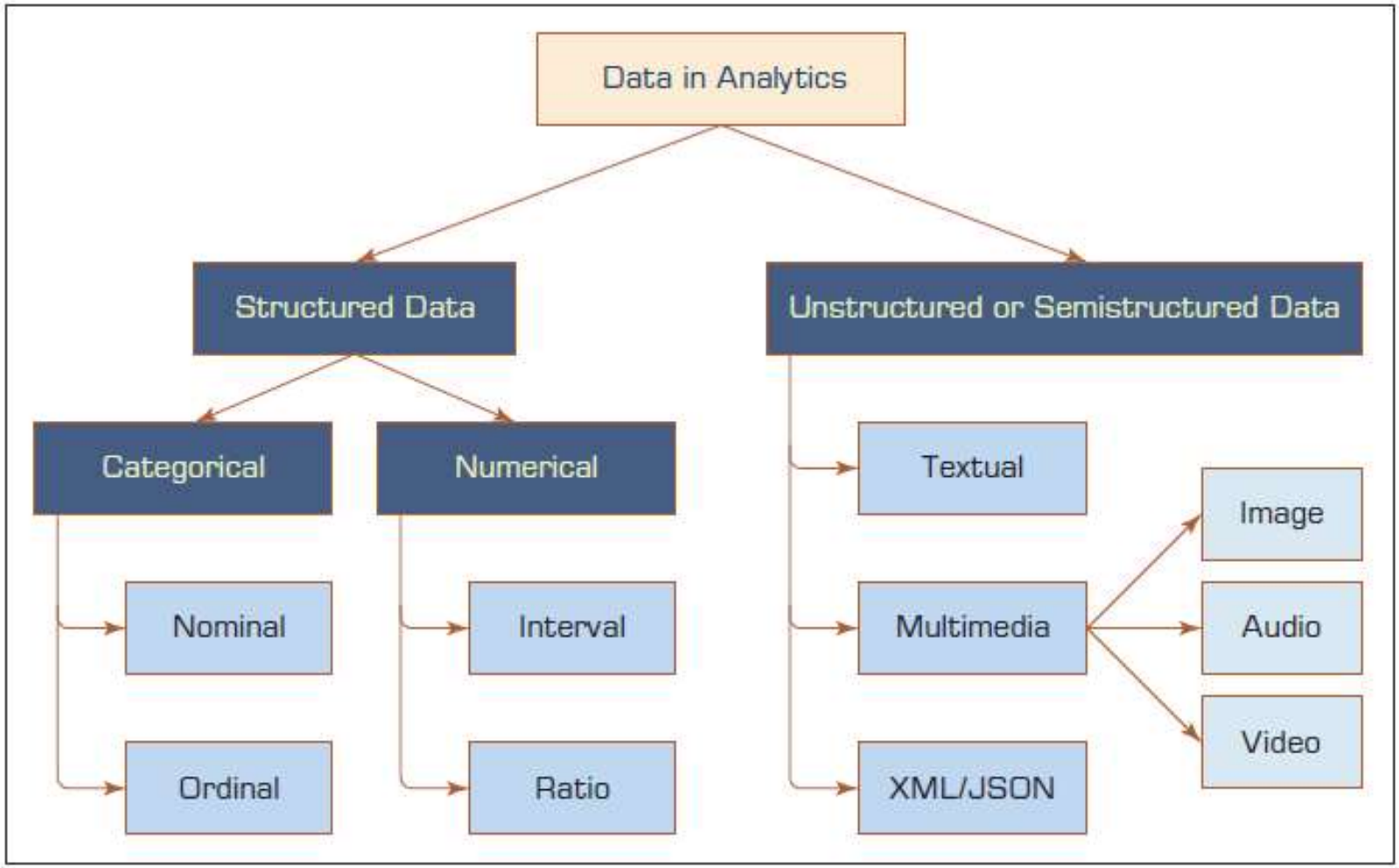
The Nature of Data



A Simple Taxonomy of Data

- Data (datum—singular form of data): facts
- Structured data
  - Targeted for computers to process
  - Numeric versus nominal
- Unstructured/textual data
  - Targeted for humans to process/digest
- Semi-structured data?
  - XML, HTML, Log files, etc.
- Data taxonomy…

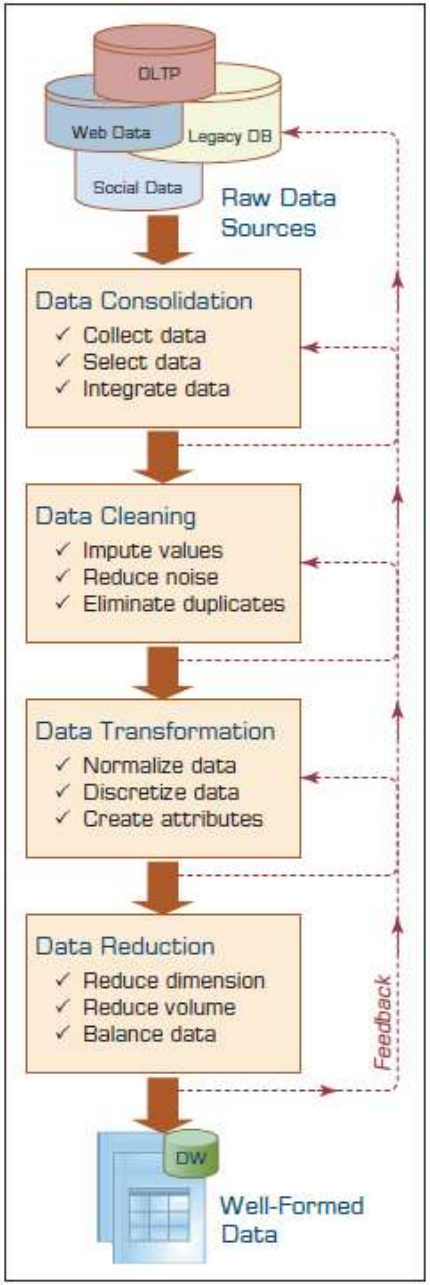
A Simple Taxonomy of Data



The Art and Science of Data Preprocessing

- The real-world data is dirty, misaligned, overly complex, and inaccurate
  - Not ready for analytics!
- Readying the data for analytics is needed
  - Data preprocessing
    - Data consolidation
    - Data cleaning
    - Data transformation
    - Data reduction
- Art – it develops and improves with experience

The Art and Science of Data Preprocessing

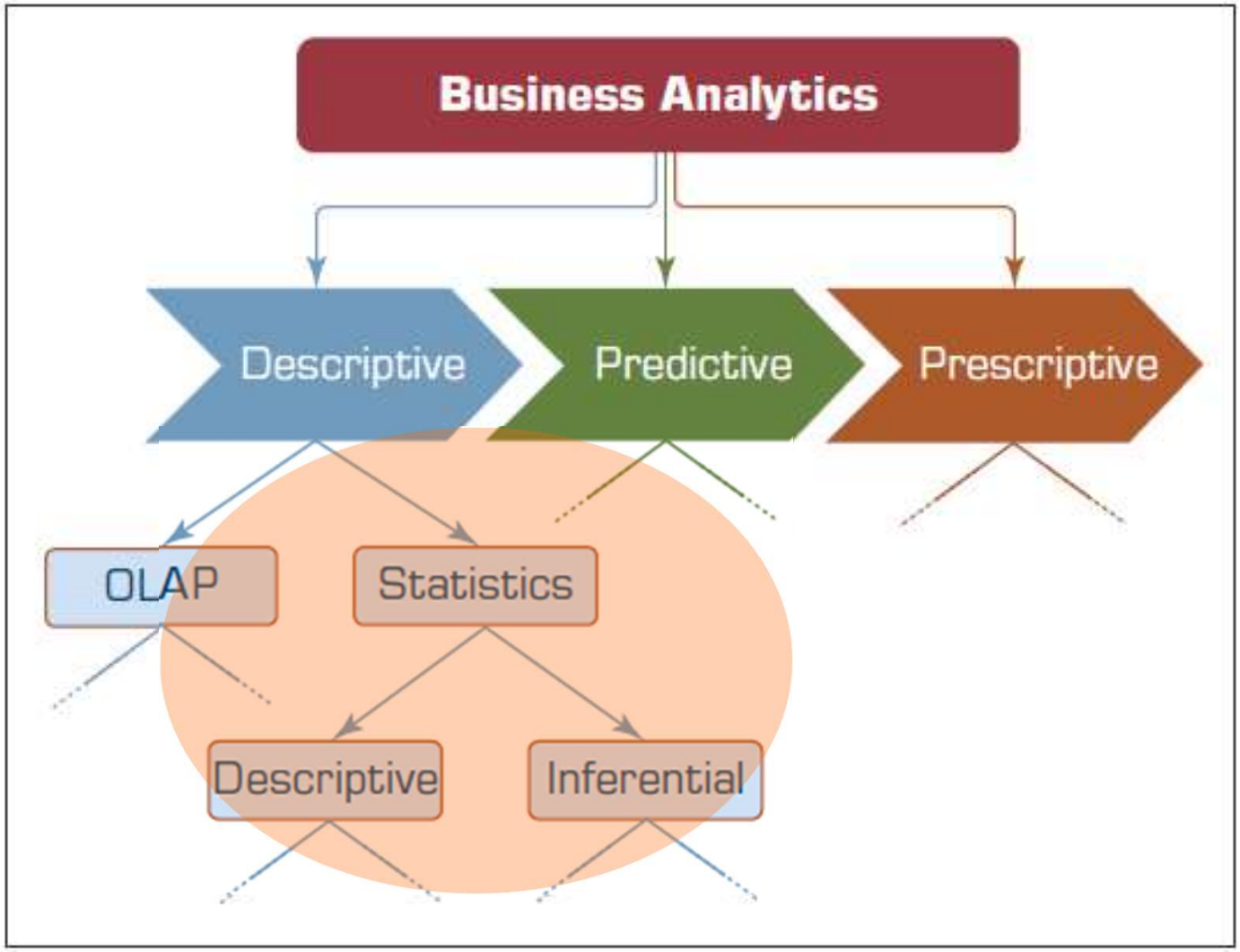


Data Preprocessing Tasks and Methods

TABLE 2.1 A Summary of Data Preprocessing Tasks and Potential Methods

Main Task	Subtasks	Popular Methods
Data consolidation	Access and collect the data	SQL queries, software agents, Web services.
	Select and filter the data	Domain expertise, SQL queries, statistical tests.
	Integrate and unify the data	SQL queries, domain expertise, ontology-driven data mapping.
Data cleaning	Handle missing values in the data	Fill in missing values (imputations) with most appropriate values (mean, median, min/max, mode, etc.); recode the missing values with a constant such as "ML"; remove the record of the missing value; do nothing.
	Identify and reduce noise in the data	Identify the outliers in data with simple statistical techniques (such as averages and standard deviations) or with cluster analysis; once identified, either remove the outliers or smooth them by using binning, regression, or simple averages.
	Find and eliminate erroneous data	Identify the erroneous values in data (other than outliers), such as odd values, inconsistent class labels, odd distributions; once identified, use domain expertise to correct the values or remove the records holding the erroneous values.
Data transformation	Normalize the data	Reduce the range of values in each numerically valued variable to a standard range (e.g., 0 to 1 or −1 to +1) by using a variety of normalization or scaling techniques.
	Discretize or aggregate the data	If needed, convert the numeric variables into discrete representations using range- or frequency-based binning techniques; for categorical variables, reduce the number of values by applying proper concept hierarchies.
	Construct new attributes	Derive new and more informative variables from the existing ones using a wide range of mathematical functions (as simple as addition and multiplication or as complex as a hybrid combination of log transformations).
Data reduction	Reduce number of attributes	Principal component analysis, independent component analysis, chi-square testing, correlation analysis, and decision tree induction.
	Reduce number of records	Random sampling, stratified sampling, expert-knowledge-driven purposeful sampling.
	Balance skewed data	Oversample the less represented or undersample the more represented classes.

Statistical Modeling for Business Analytics

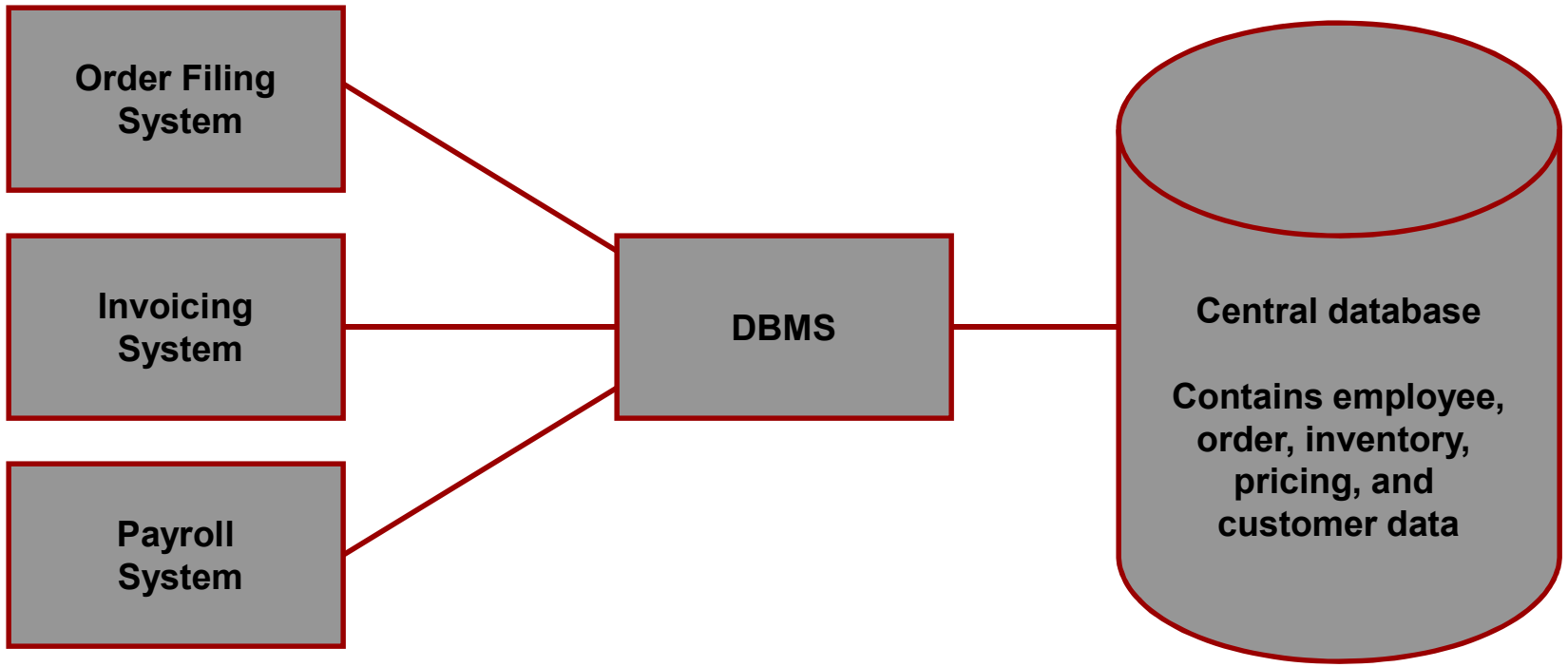


Data

- Data need to be processed
  1. Structured
     - Employee record, Product, Order, Transaction, etc.
  2. Unstructured/Semistructured
     - Email, Tweets, blogs, social chats, reviews, etc.

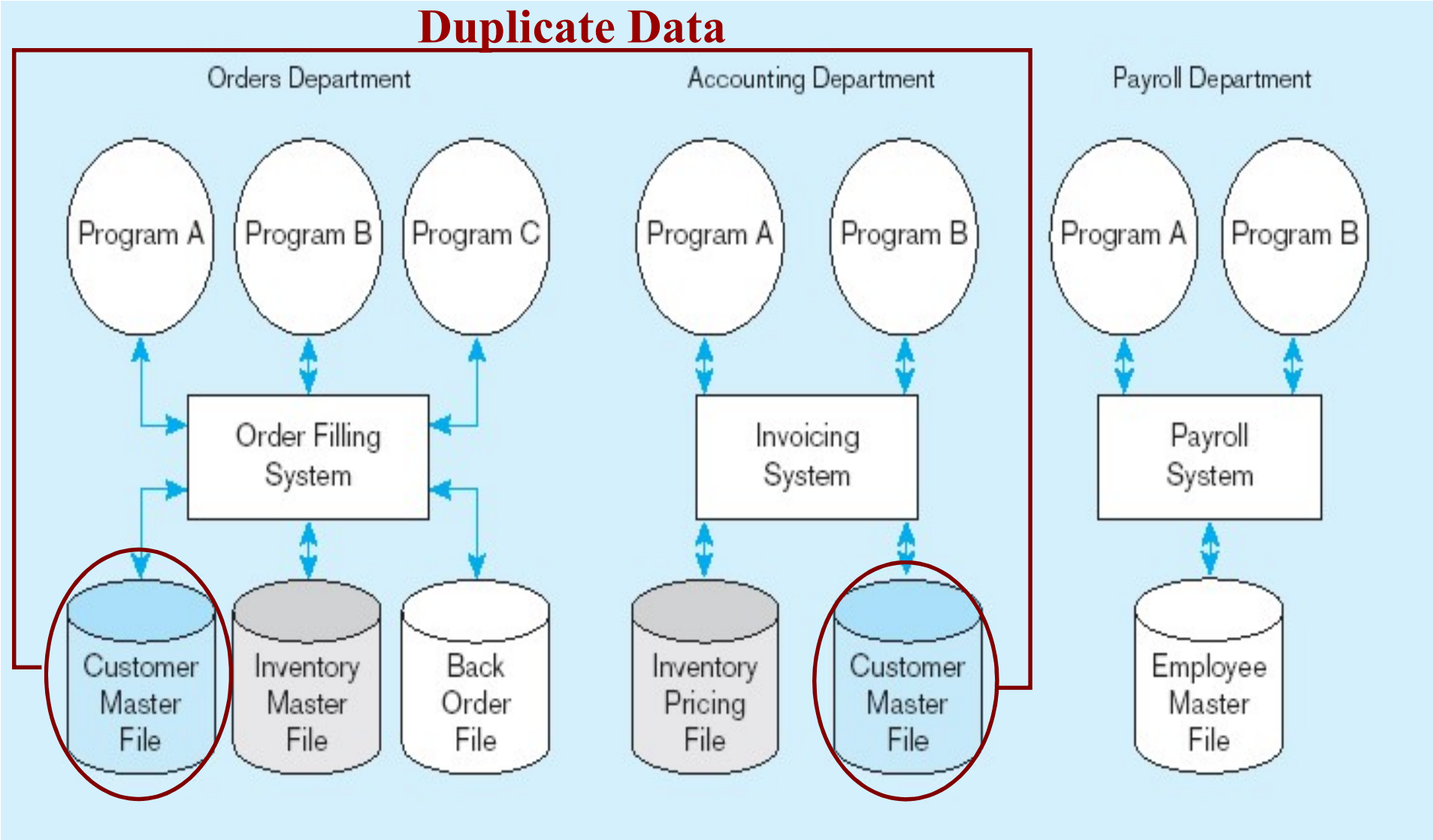
Database Management System

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



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

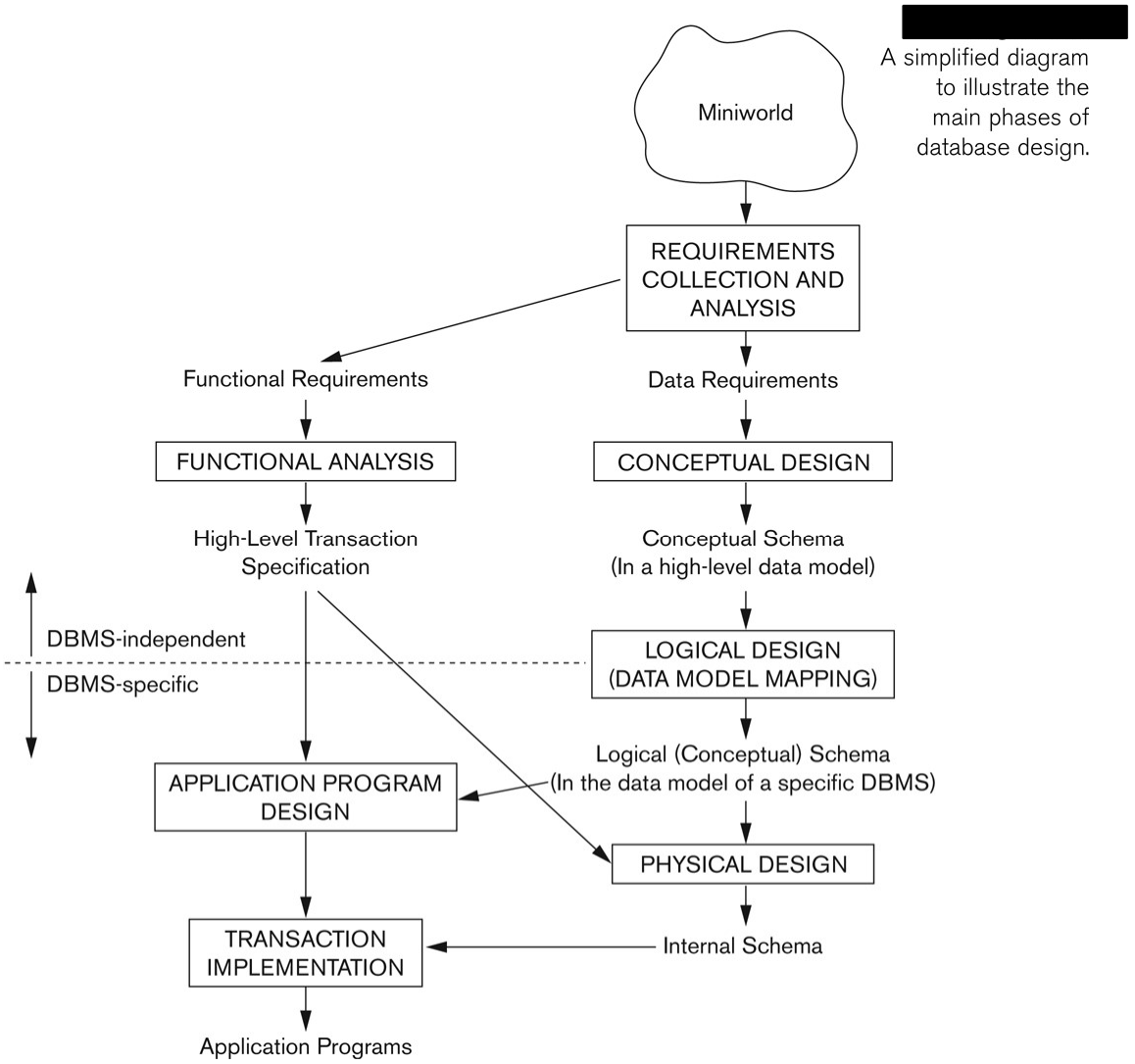
- Program-data independence
- Planned data redundancy
- Improved data consistency
- Improved data sharing
- Increased application development productivity
- Enforcement of standards
- Improved data quality
- Improved data accessibility and responsiveness
- Reduced program maintenance
- Improved decision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

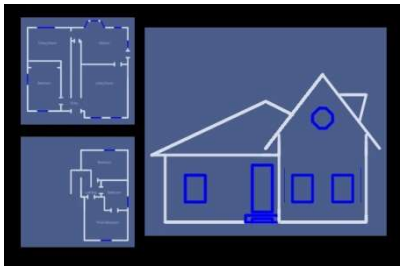
Overview of Database Design Process



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

How to build Database Application?

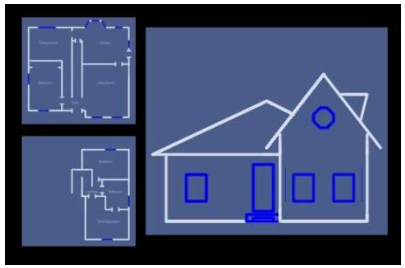
1. Blue print
2. Tools
3. Construction
4. House




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

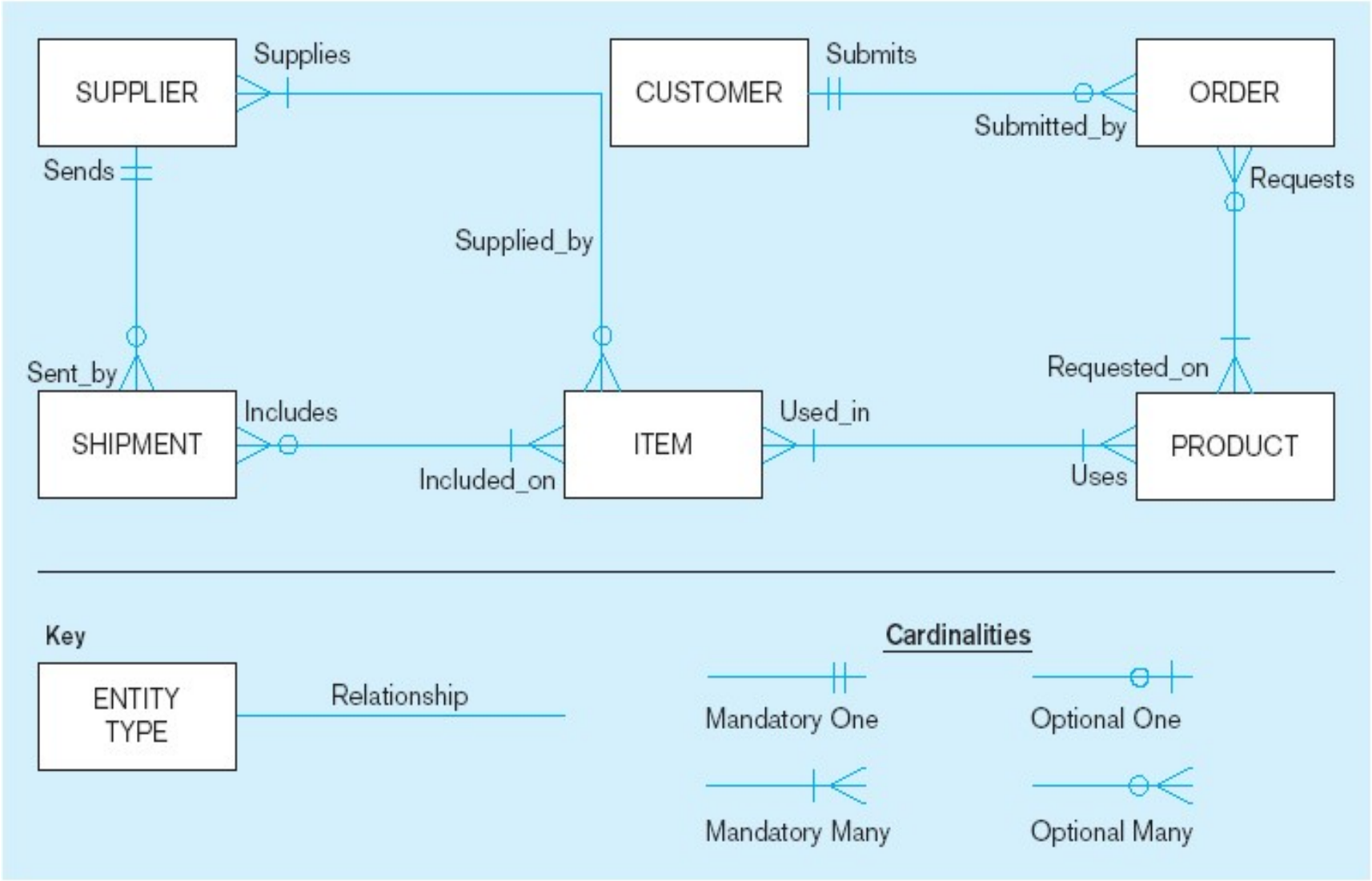
How to build Database Application?

2. Relations
   - Normalization
   - Normal Forms
     - 1st Normal Form
     - 2nd Normal Form
     - 3rd Normal Form
     - BCNF

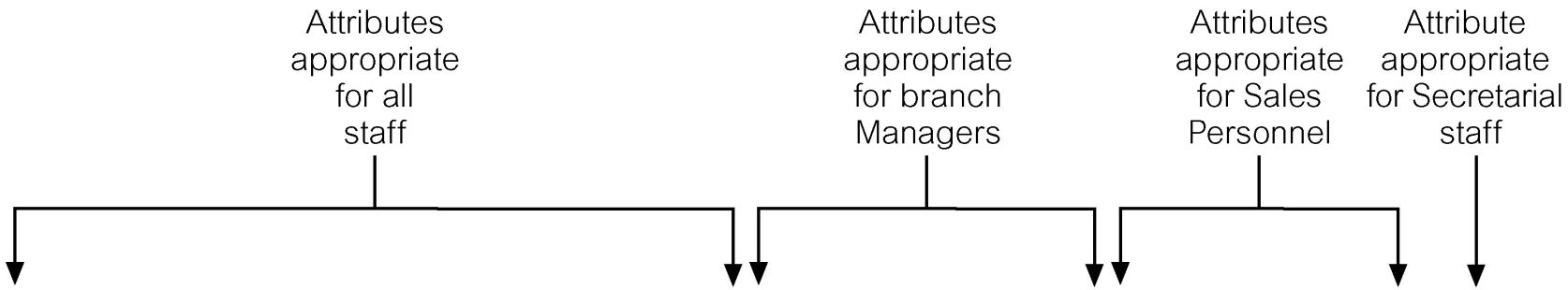
How to build Database Application?

4. Tables
   - Rows
   - Columns
   - Primary keys
   - Foreign Keys
   - Constraints

Sample E-R Diagram

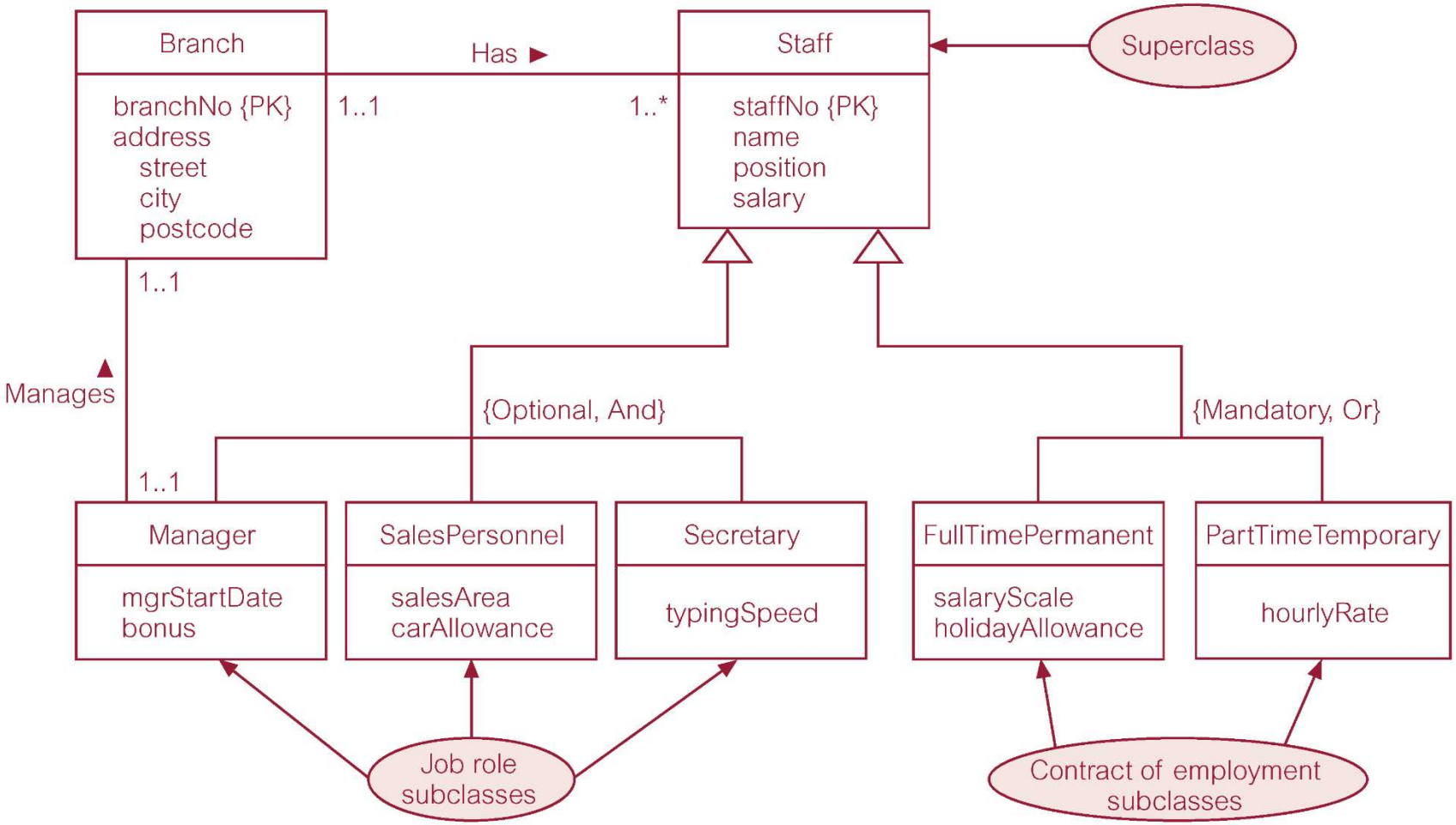


AllStaff relation holding details of all staff



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



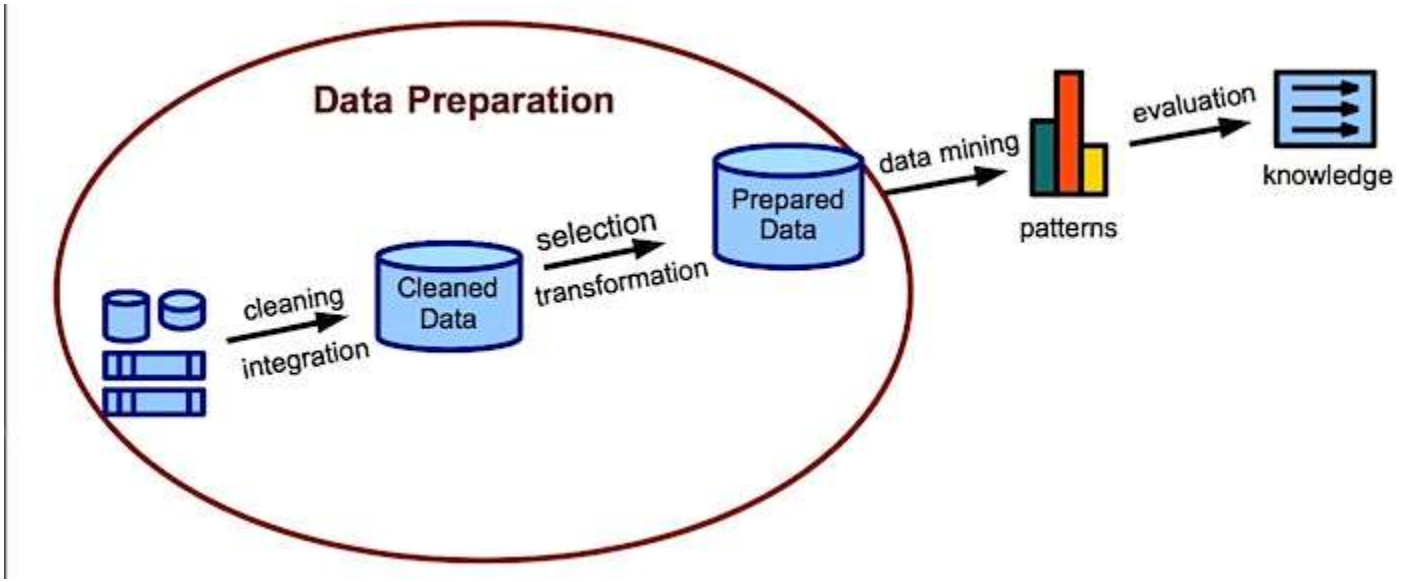
Data , what is in the name?

- 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

REFERENCE : Social Media Analytics: Data Mining Applied to Insurance Twitter Posts Roosevelt C. Mosley Jr., FCAS, MAAA

Big Data Prime Time

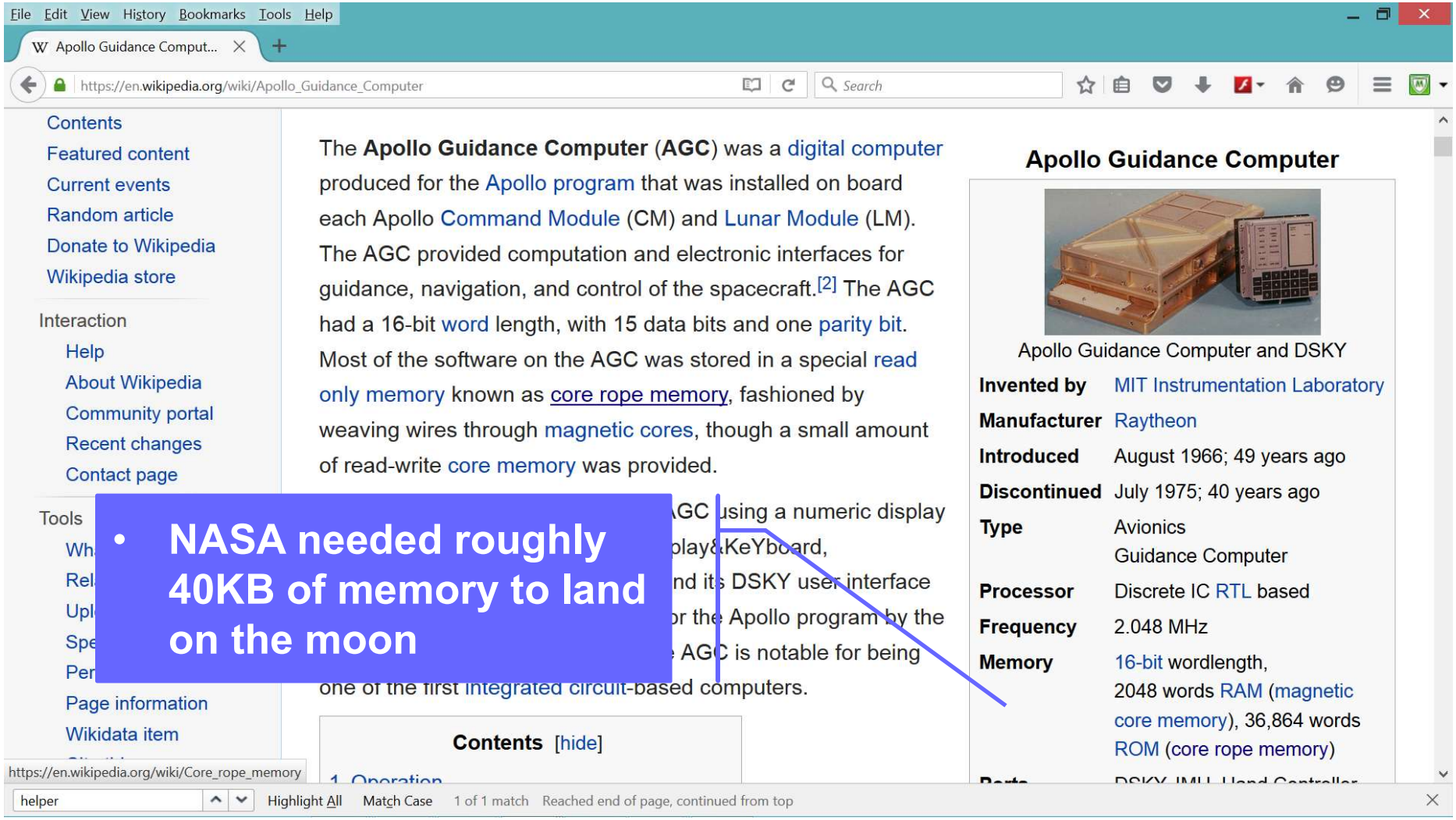
How much data you think roughly out there ?

- 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.

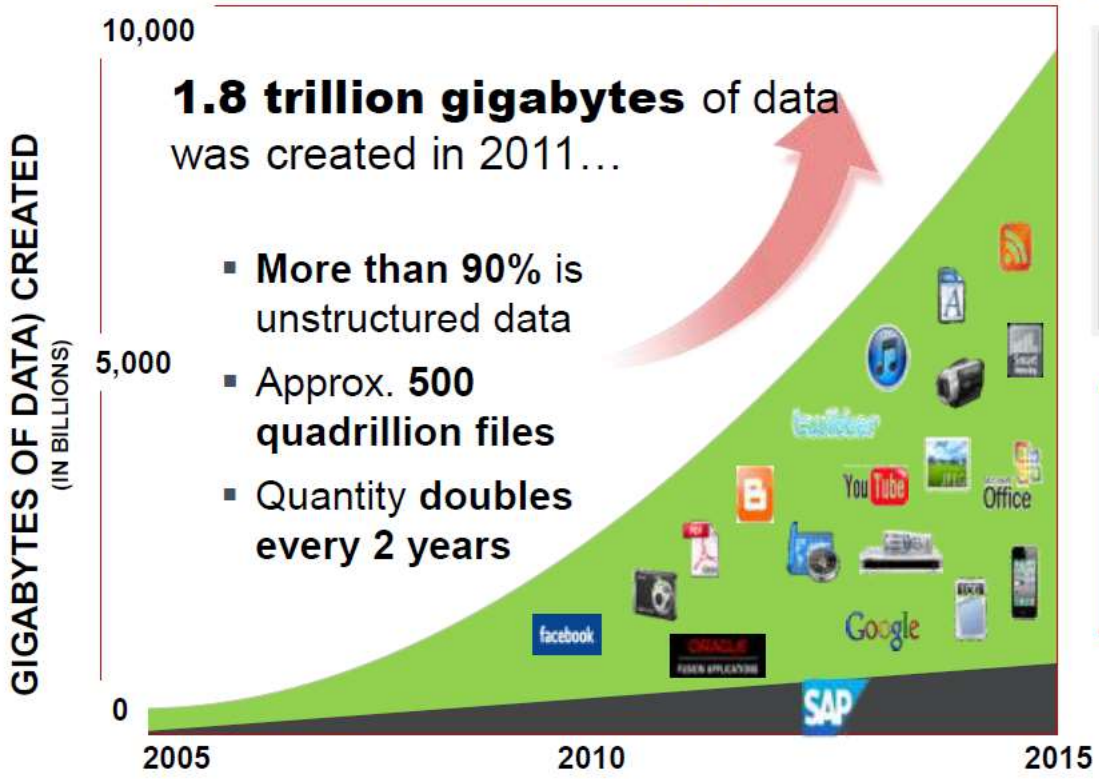
Big Data Prime Time



- **NASA needed roughly 40KB of memory to land on the moon**

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."

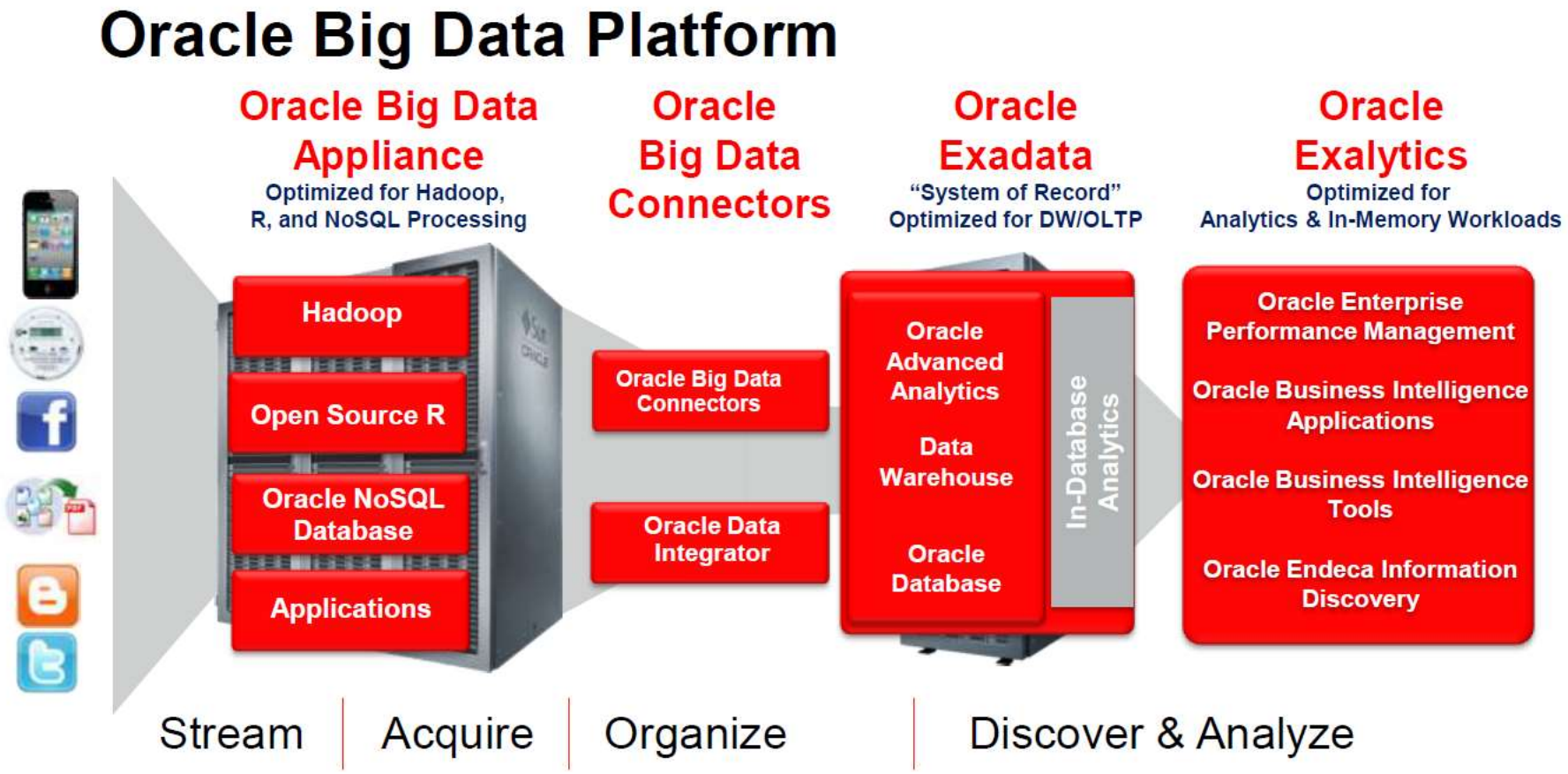
- Google CEO Eric Schmidt

Requires capability to rapidly:

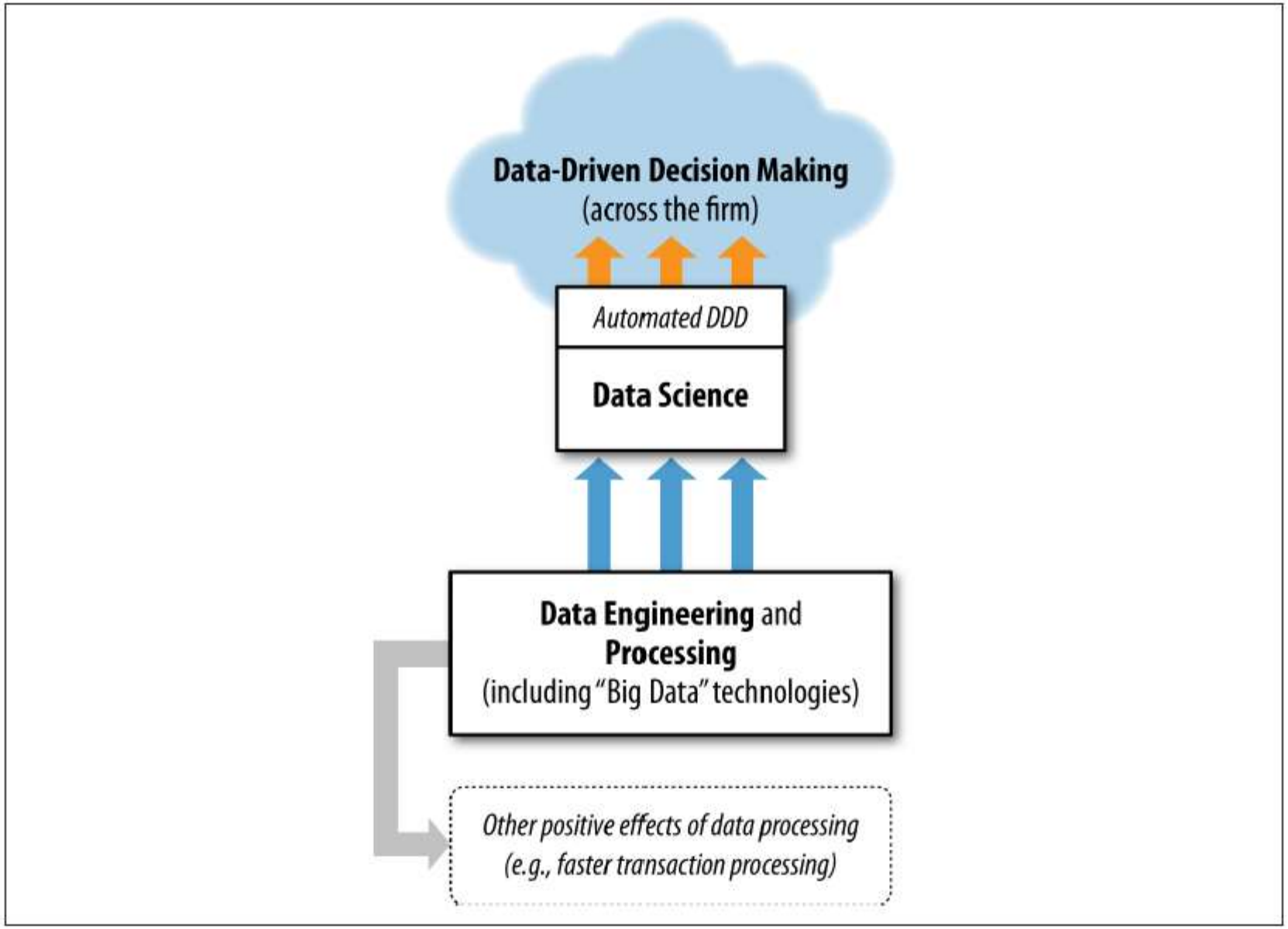
- ✓ Collect and integrate data
- ✓ ***Understand data & their relationships***
- ✓ ***Respond and take action***

Big Data Prime Time

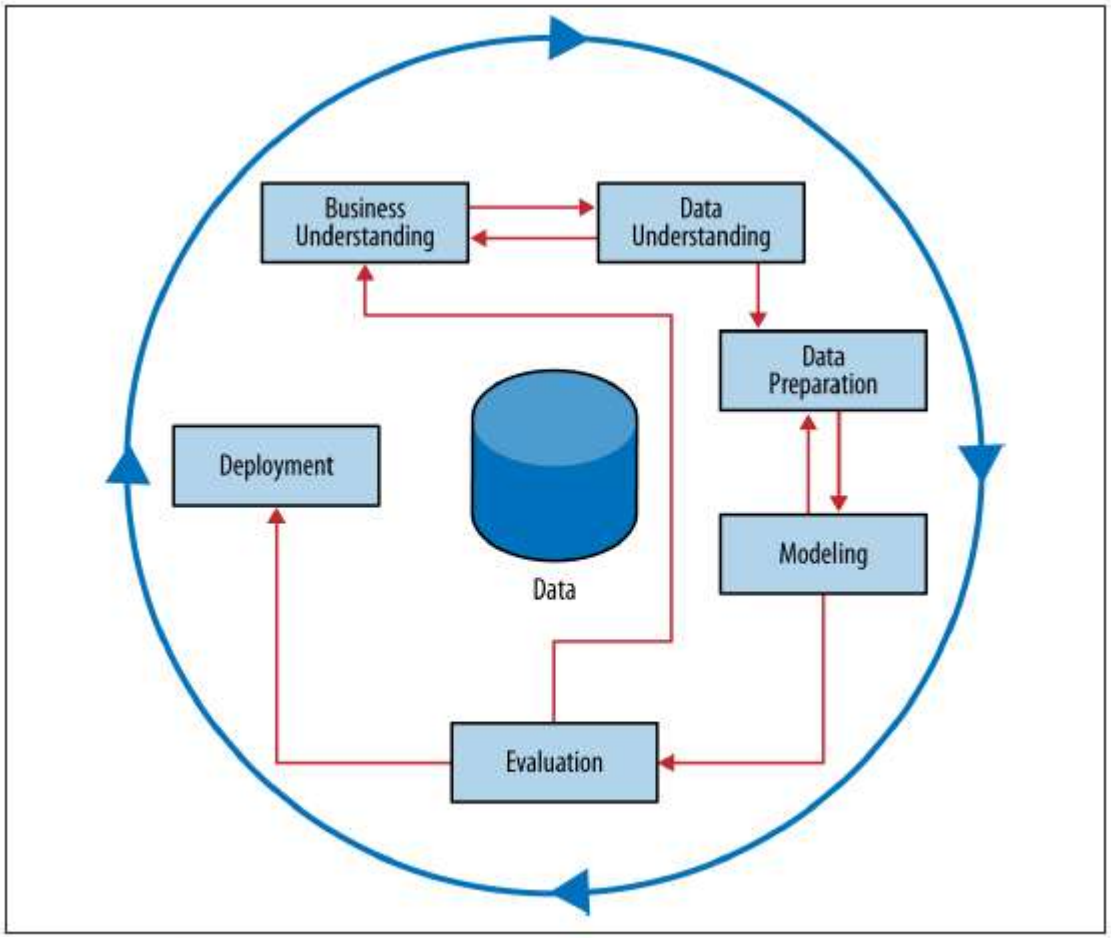
Is RDBMS and Structured data history by now? No …



The Process

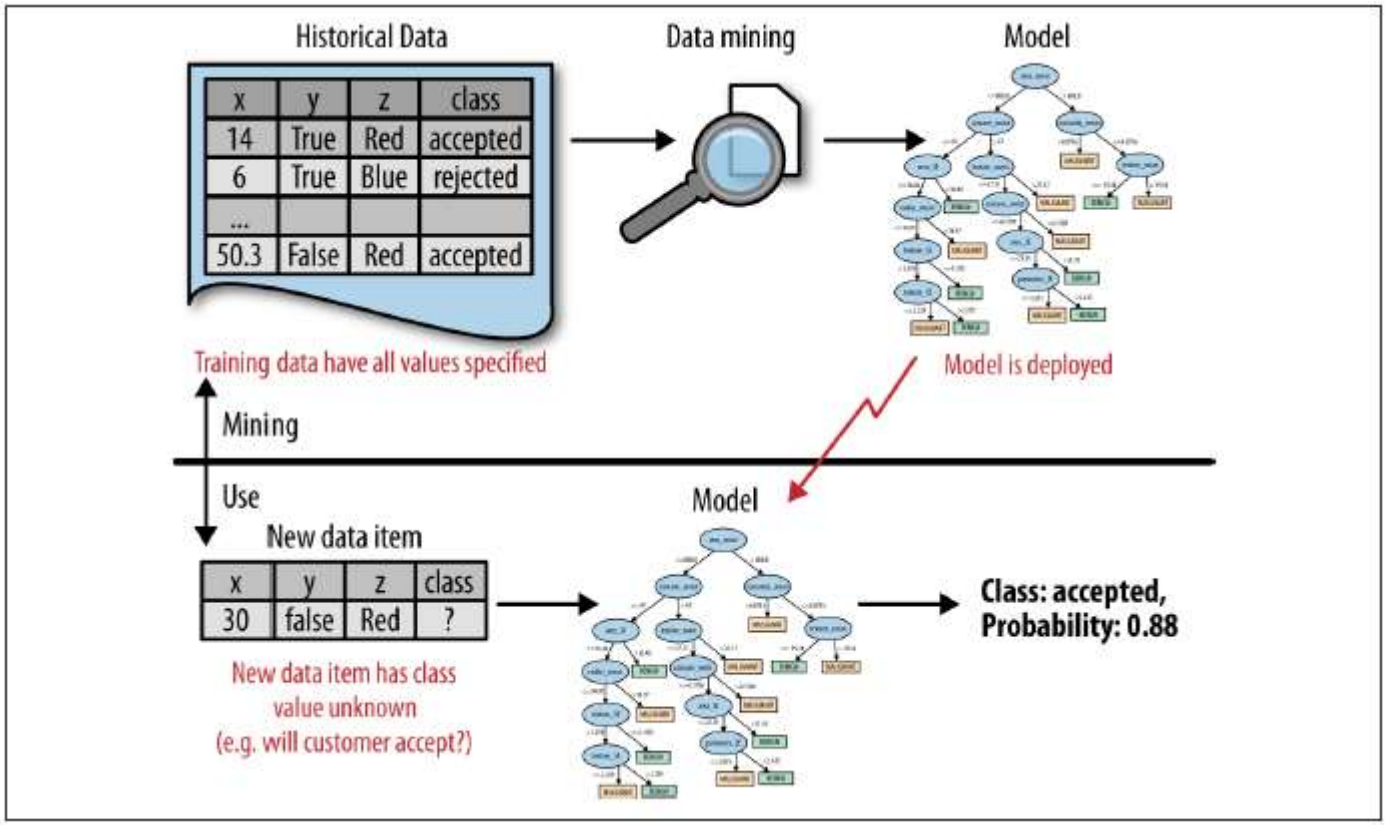

Data-Driven Decision Making
(across the firm)
Automated DDD
Data Science
Data Engineering and
Processing
(including "Big Data" technologies)
Other positive effects of data processing
(e.g., faster transaction processing)

The Process



The CRISP data mining process.

The Process


Historical Data
Data mining
Model
x
y
z
class
14
True
Red
accepted
6
True
Blue
rejected
...
50.3
False
Red
accepted
Training data have all values specified
Model is deployed
Mining
Use
Model
New data item
x
y
z
class
30
false
Red
?
New data item has class value unknown (e.g. will customer accept?)
Class: accepted, Probability: 0.88

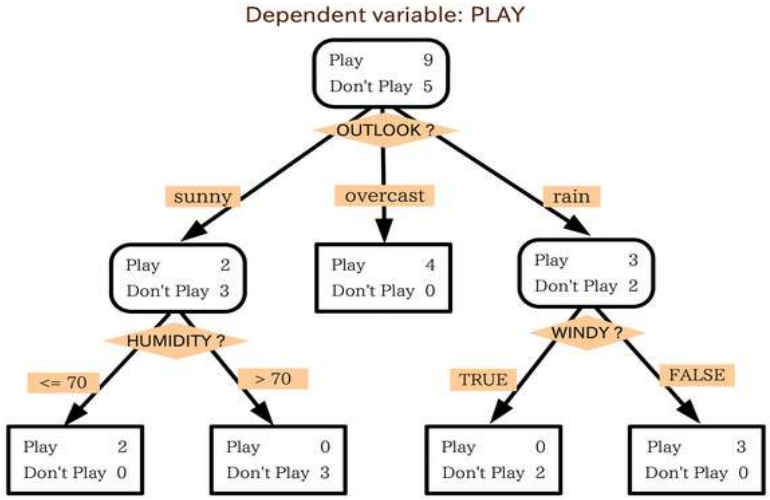
Data Format

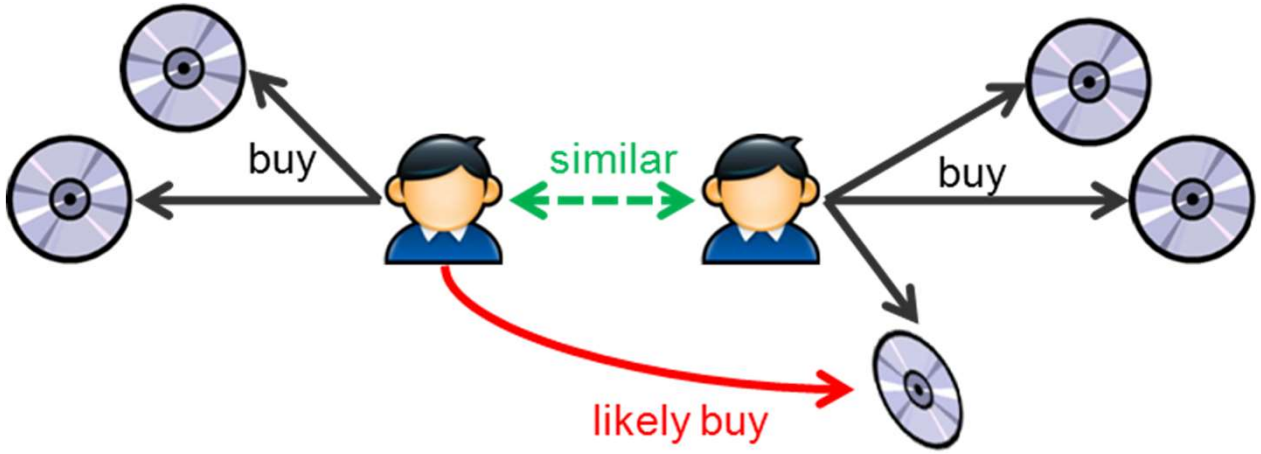
- Data is represented by ASCII code (American Standard Code for Information Interchange) which is the most widely used format.

- 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).

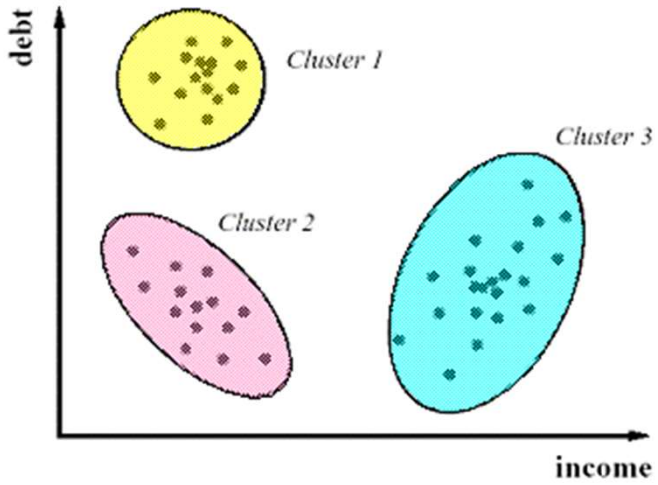
  2. Binary. Data is structured as a record by fixed blocks (formatted text)

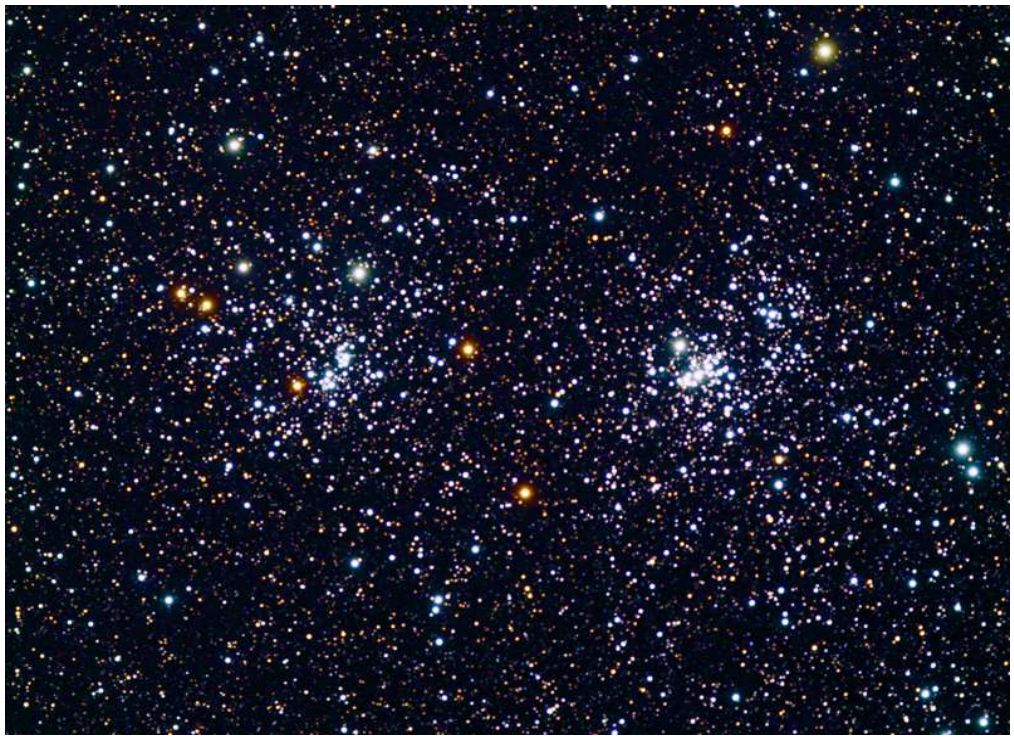
Structured and Unstructured Data …



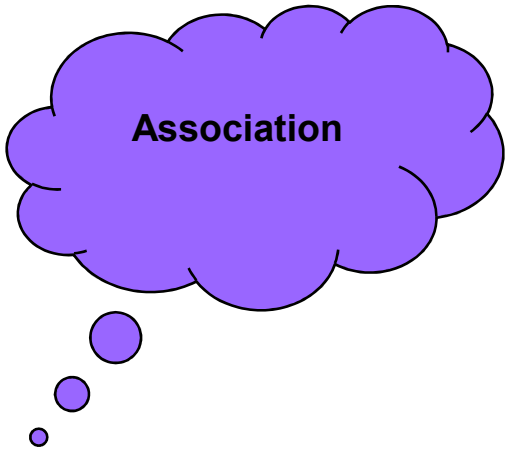


Structured and Unstructured Data …

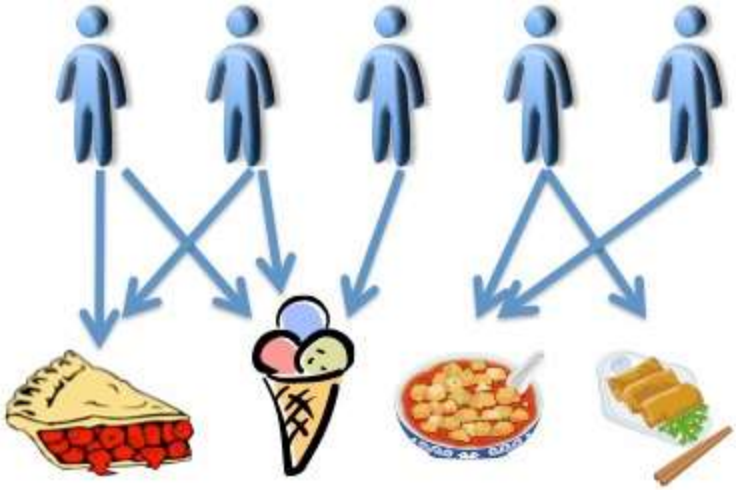


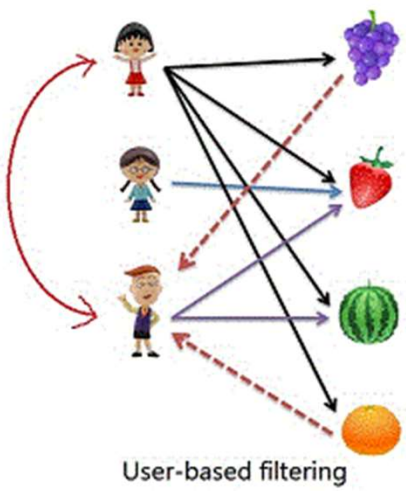


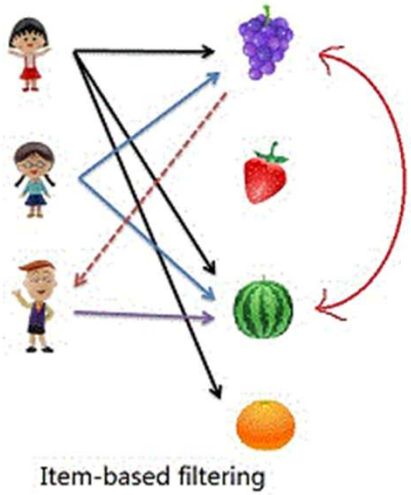
Structured and Unstructured Data …



Structured and Unstructured Data …







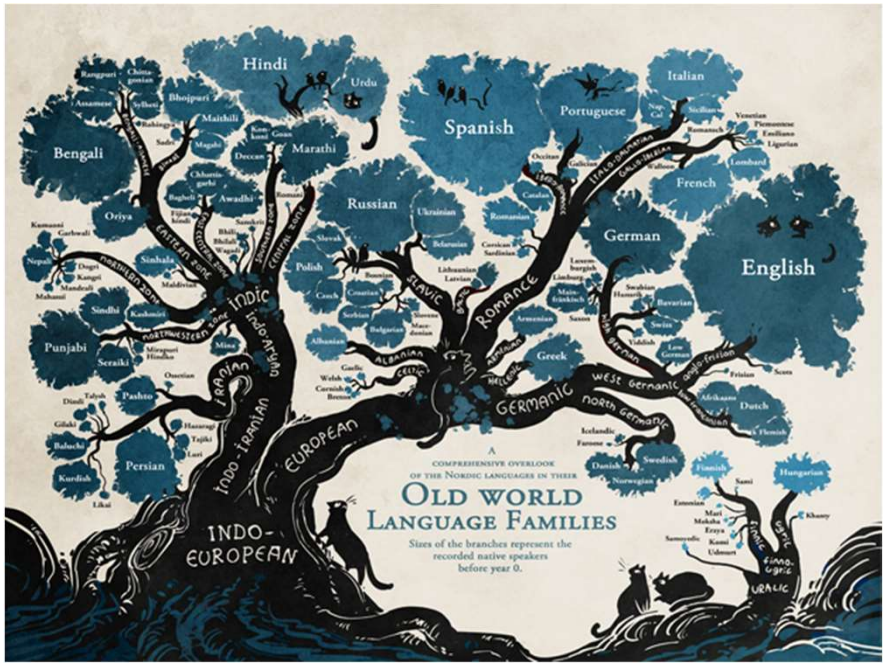
Structured and Unstructured Data …

Technologies and Tools

- So, what are the technologies and tools that we will use in this class?

Technologies and Tools

- How many Natural Languages out there?
  - **There are 6,909 distinct languages**
- How many Programming Languages out there?
  - More than a thousand invented over the years … But how many of them in use as of today in the software industry?
  - The first commercial supported high-level programing language is FORTRAN 1956



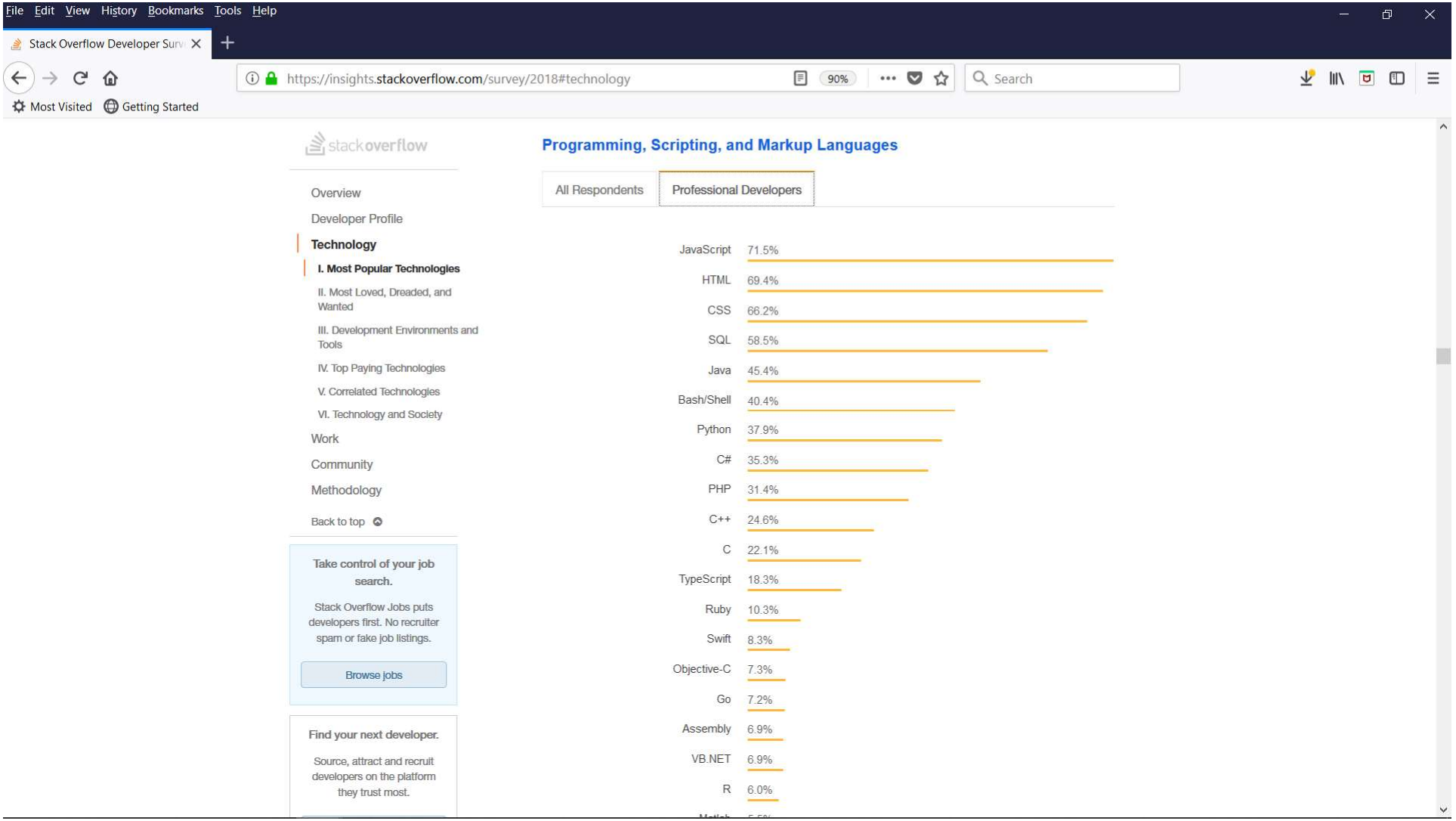
Technologies and Tools

- What does it mean high-level programming language?
  - **C, C++, Java, etc.**
- What does it mean scripting programming language?
  - **KSH, Bash, JavaScript, Python, etc.**
- What does it mean query language?
  - **SQL and NoSQL**
- Scripting language can refer to dynamic high-level general-purpose languages
  - **Perl, Python, TCL**
- Interpreted vs. Compiled programming languages?

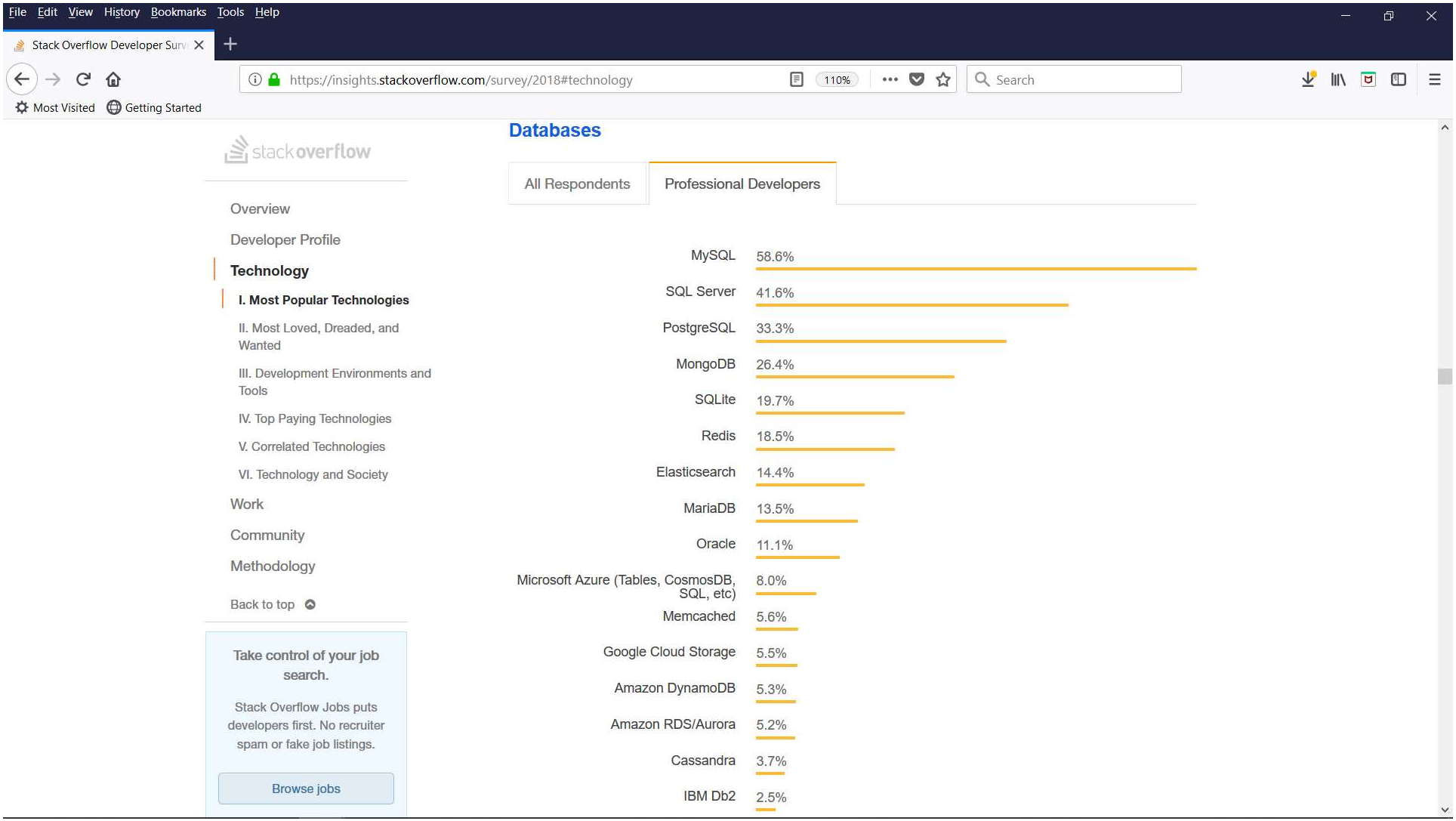
Timeline for Programming Languages

- 1950-1959
  - **Fortran, LISP, COBOL**
- 1960-1969
  - **ALGOL 68, BASIC, Simula**
- 1970-1979
  - **Pascal, C, SmalTalk, AWK, MATLAB, SAS**
- 1980-1989
  - **Ada, C++, Objective-C, PostScript, TCL**
- 1990-1999
  - **VisualBasic, Perl, Python, R, Java, JavaScript, PHP, Ruby**
- 2000-2009
  - **C#, Scala, Go (Google), Typescript (Microsoft)**
- 2010-
  - **Typescript, Swift (Apple)**

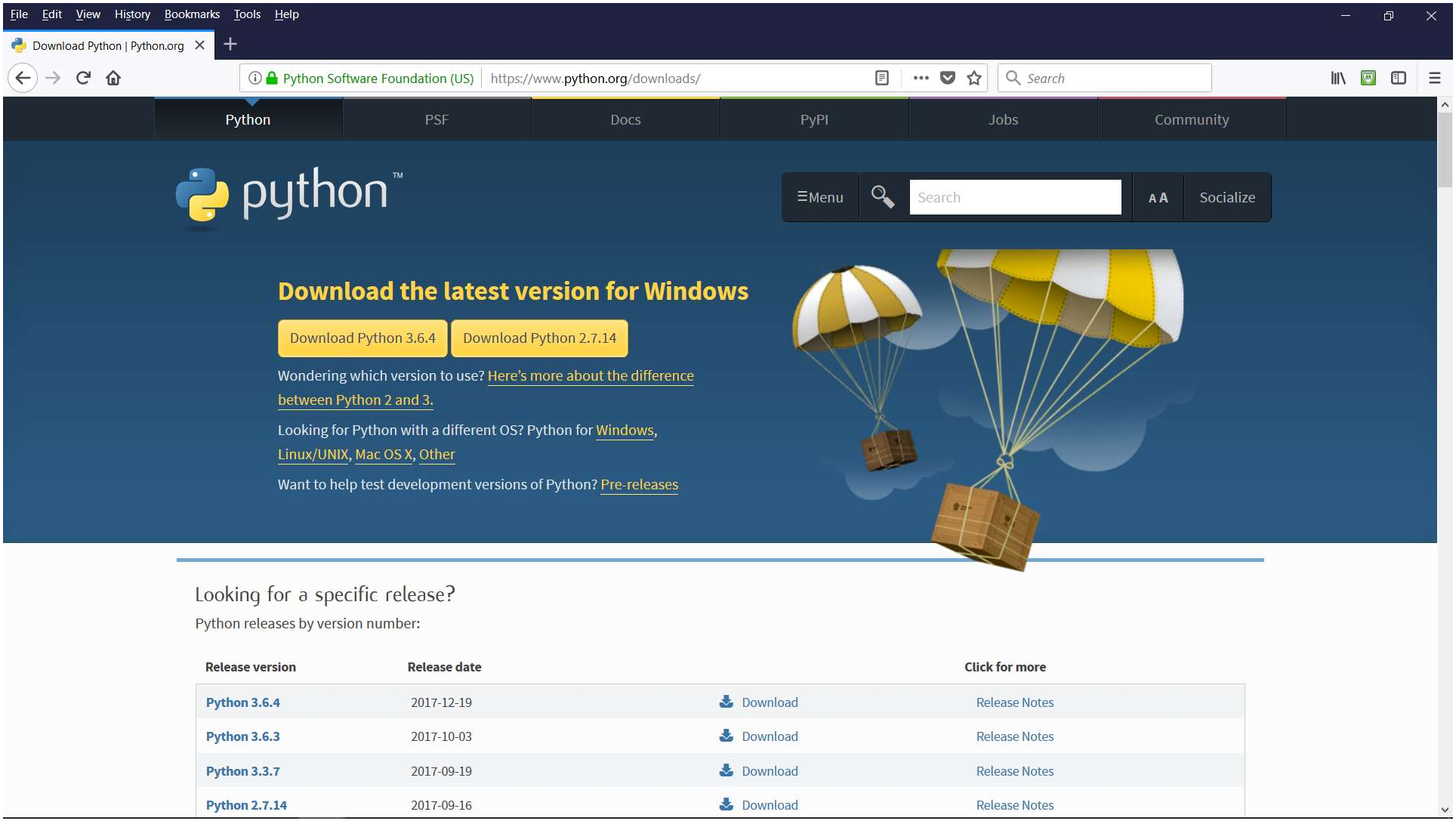
Technologies and Tools



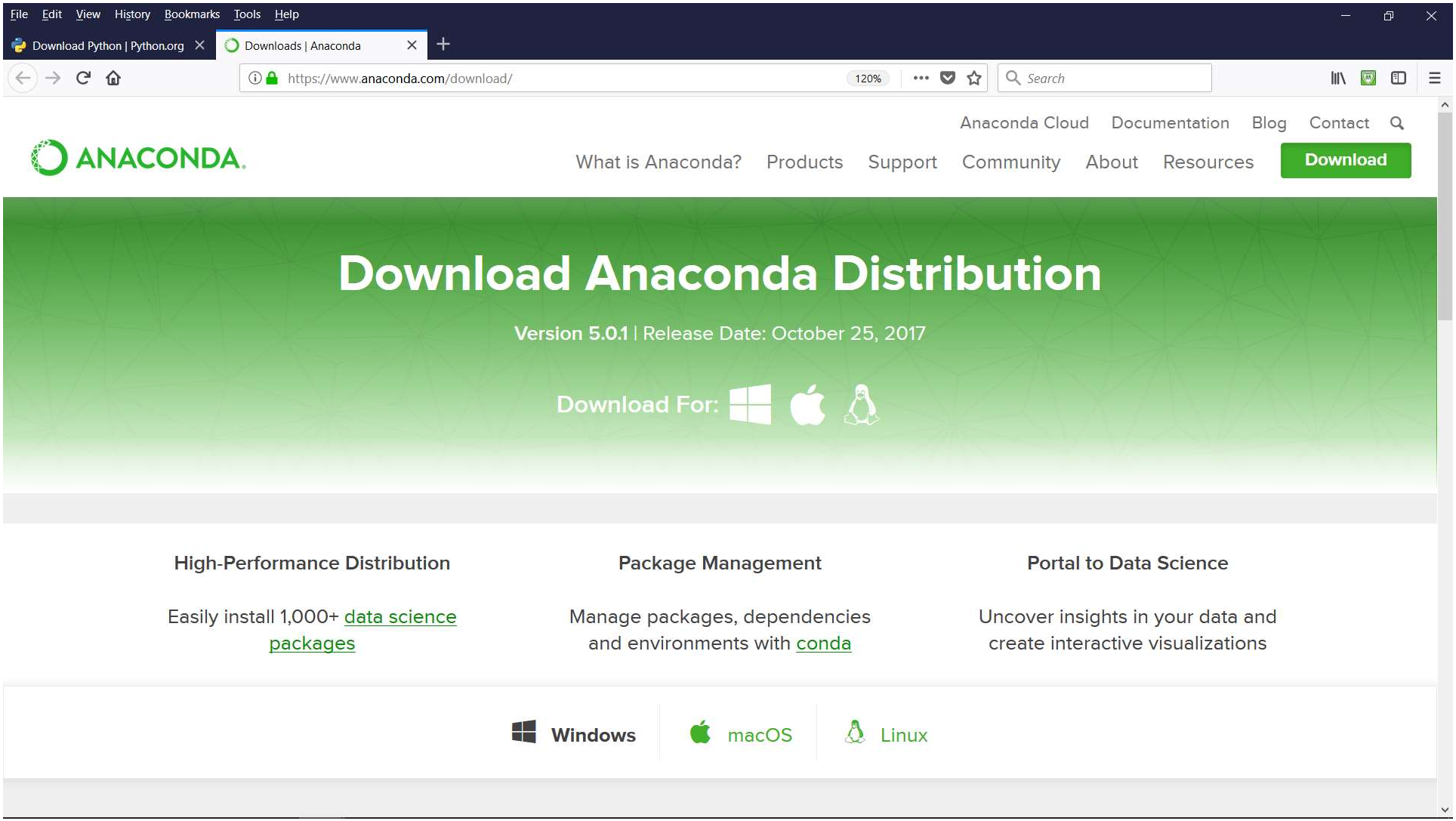
Technologies and Tools



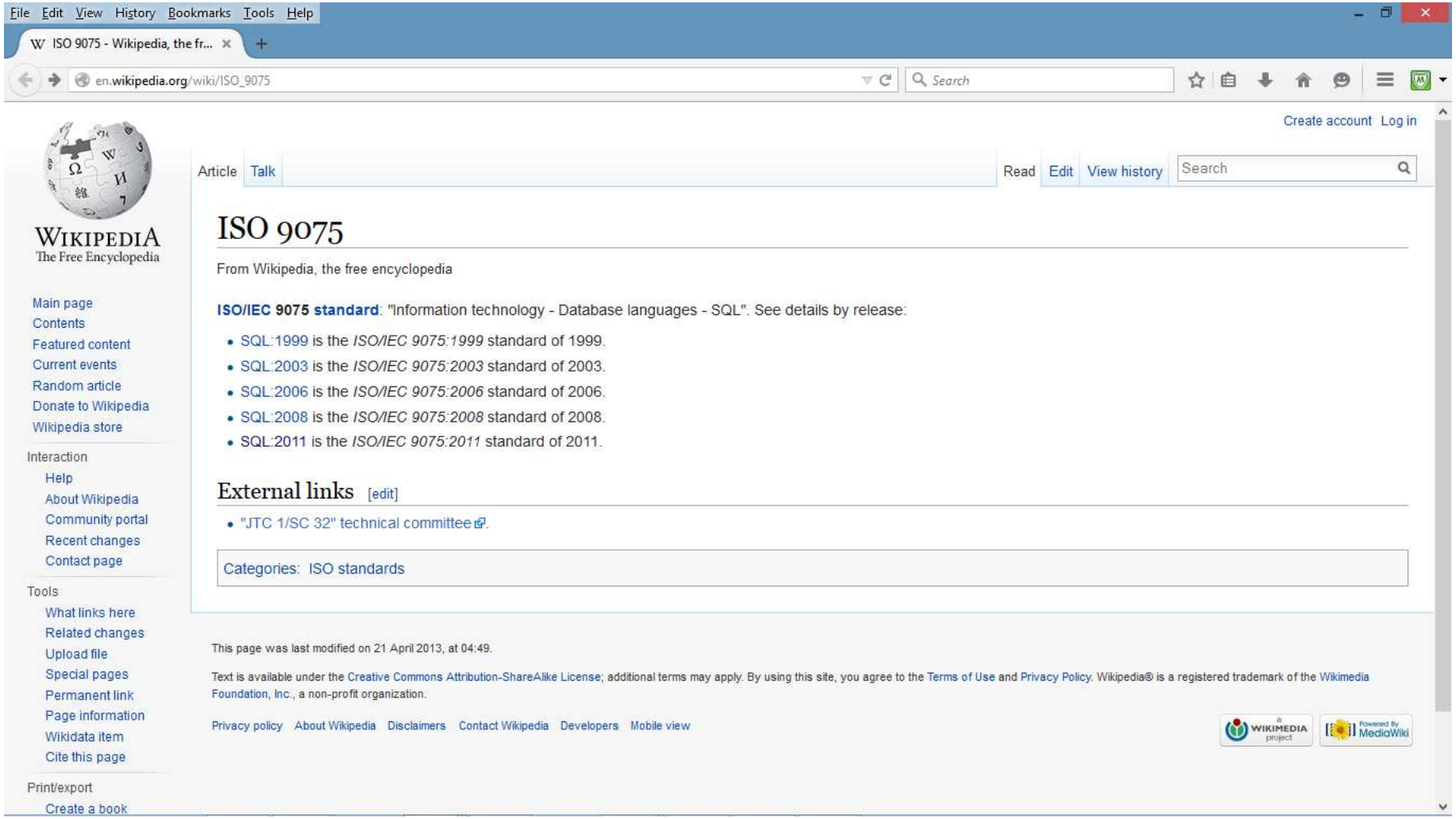
Python


File Edit View History Bookmarks Tools Help
Download Python | Python.org
Python Software Foundation (US)
https://www.python.org/downloads/
Search
Python
PSF
Docs
PyPI
Jobs
Community
python™
☰Menu
Search
A A
Socialize
Download the latest version for Windows
Download Python 3.6.4
Download Python 2.7.14
Wondering which version to use? Here's more about the difference between Python 2 and 3.
Looking for Python with a different OS? Python for Windows, Linux/UNIX, Mac OS X, Other
Want to help test development versions of Python? Pre-releases
Looking for a specific release?
Python releases by version number:
Release version
Release date
Click for more
Python 3.6.4 2017-12-19 Download Release Notes
Python 3.6.3 2017-10-03 Download Release Notes
Python 3.3.7 2017-09-19 Download Release Notes
Python 2.7.14 2017-09-16 Download Release Notes

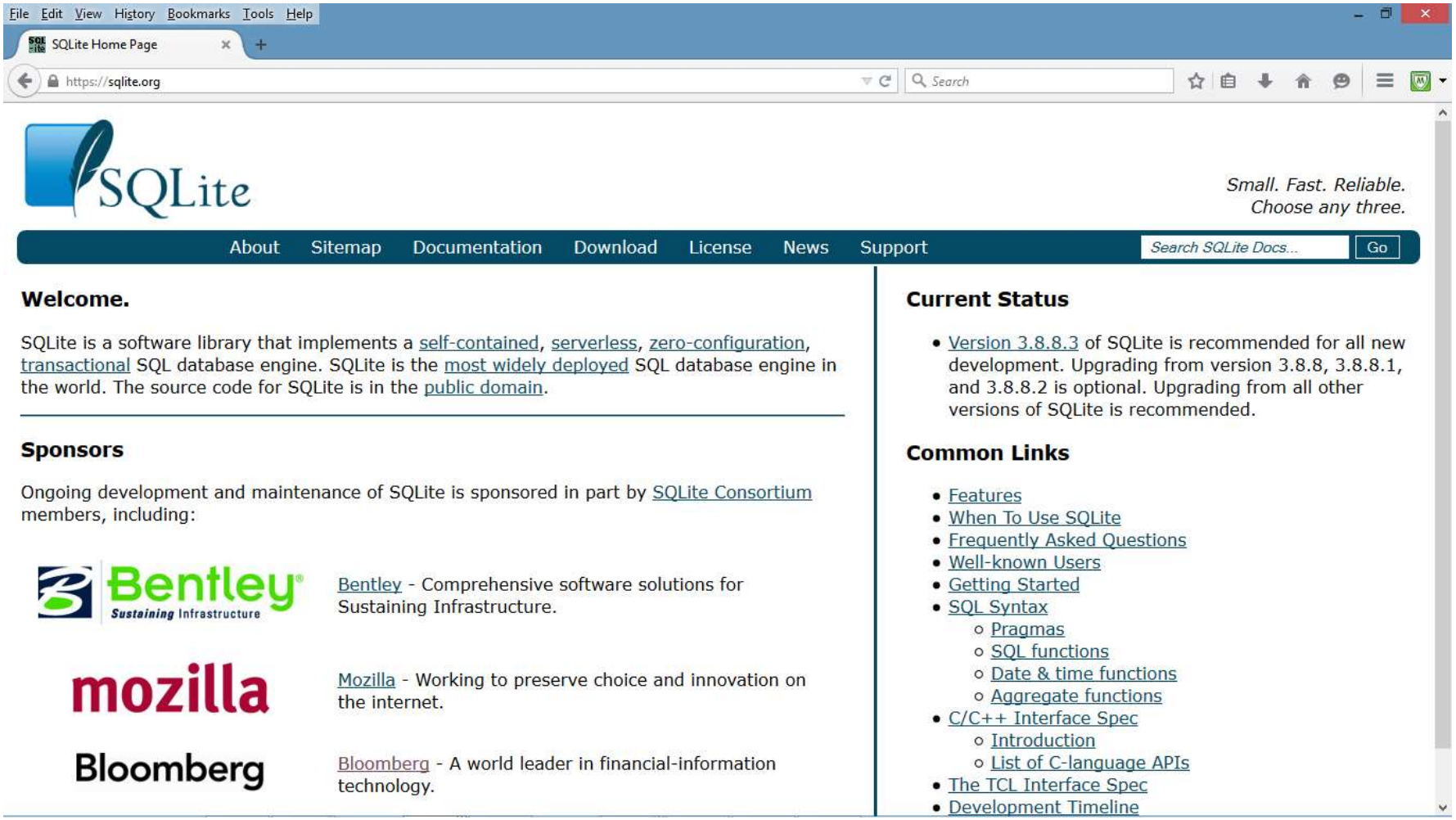
Anaconda


File Edit View History Bookmarks Tools Help
Download Python | Python.org
Downloads | Anaconda
https://www.anaconda.com/download/
120%
Search
Anaconda Cloud Documentation Blog Contact
ANACONDA.
What is Anaconda? Products Support Community About Resources
Download
Download Anaconda Distribution
Version 5.0.1 | Release Date: October 25, 2017
Download For:
High-Performance Distribution
Easily install 1,000+ data science packages
Package Management
Manage packages, dependencies and environments with conda
Portal to Data Science
Uncover insights in your data and create interactive visualizations
Windows
macOS
Linux

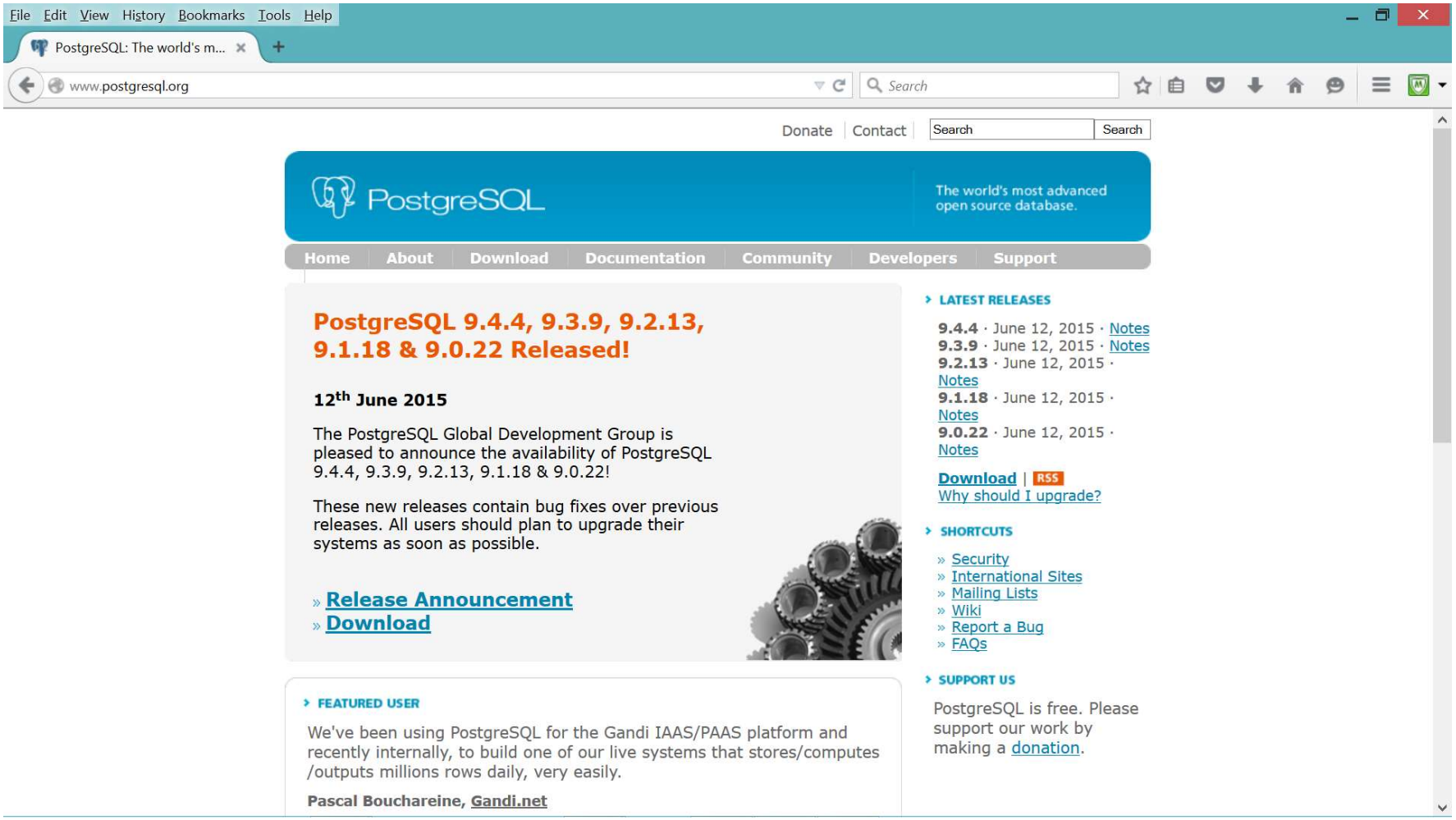
SQL


File Edit View History Bookmarks Tools Help
ISO 9075 - Wikipedia, the fr...
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

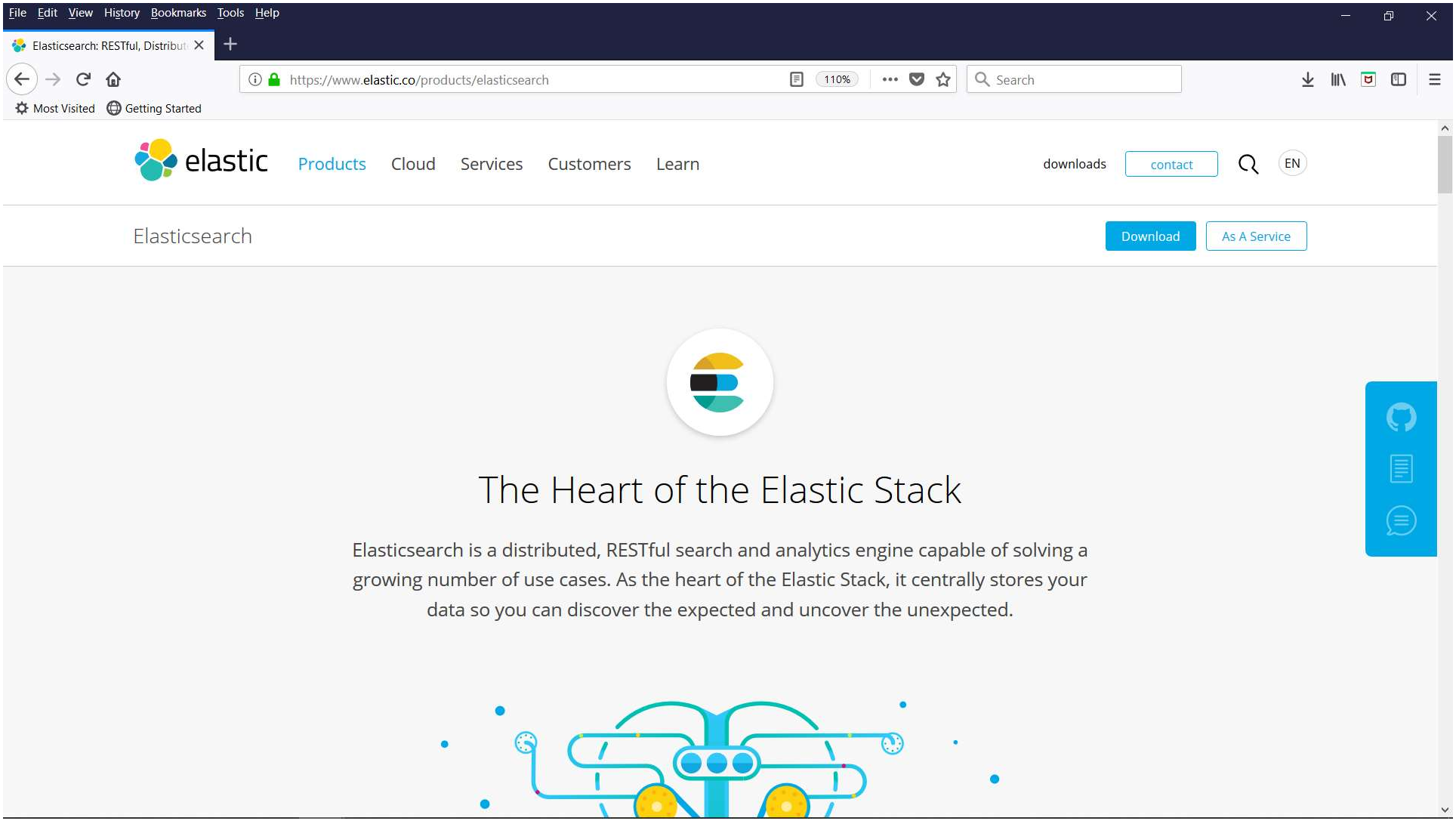
SQLite


File Edit View History Bookmarks Tools Help
SQLite Home Page
https://sqlite.org
Search
SQLite
Small. Fast. Reliable.
Choose any three.
About Sitemap Documentation Download License News Support
Search SQLite Docs...
Go
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
Bentley - Comprehensive software solutions for Sustaining Infrastructure.
mozilla
Mozilla - Working to preserve choice and innovation on the internet.
Bloomberg
Bloomberg - A world leader in financial-information technology.
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


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!
12th 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
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
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
SUPPORT US
PostgreSQL is free. Please support our work by making a donation.

ElasticSearch



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/NoSQL

2. What you will do in this class?
   - Weekly class discussions
   - Assignments (covering different topics)
   - Exercises (Python programming – data preparation/processing)
   - Final Exam – Proctored through Examity