

INFSCI 2710
Database Management

Week 1

Hi! My name is
Chun-Hua Tsai
(Ronald)

Today's Plan

- Introduction to course and syllabus overview
- What is data and why do we care?
- Database Management Systems overview
- Tools overview

Introduction to course and syllabus overview

What is Data and Why Do We Care?

Database Management Systems overview

Tools overview

Course Information

- **Term:** Fall 2018
- **Time:** Monday 18:00 – 20:50
- **Location:** Information Science Building, Room 403

Course Instructor

- **Instructor:** Chun-Hua Tsai
- **Email:** cht77@pitt.edu
- **Office:** Room 707, IS Building
- **Office Hours:**
 - Mon 4:30 PM – 5:30 PM
 - By appointment

Textbook

- Raghu Ramakrishnan, Johannes Gehrke. Database Management Systems. 3d Edition. WCB/McGraw-Hill 2003 (Main)
- Silberschatz, Korth, and Sudarshan Database System Concepts, 6th edition , McGraw-Hill, 2010 (Recommended).
- This book is available for free through the University of Pittsburgh library system (<http://library.pitt.edu/>).
- Direct link is <http://codex.cs.yale.edu/avi/db-book/>

You are also responsible for any information/materials

- assigned readings from the textbook
- presented during the lecture by instructor or guest speakers
- linked to from the lecture slides (any links included in the slides are fair game on the quizzes and exams)

Objectives

- Develop solid understanding of database management systems
- Understand how to design and implement relational databases
- Learn how to ask the right questions about data and how to receive (hopefully) correct answers

Objectives

- Become proficient at data modeling and writing SQL queries
- Manage administrative tasks required in a database management environment
- Learn to import and export unstructured data using PHP programming language

Grading Policy

- Homework (Lab) Assignments: 20%
- In-class activities: 10%
- Midterm Exam: 20%
- Final Exam: 25%
- Final Project: 25%

Assignments

- All of the homework assignments will be individual
- All assignments must be typed (handwritten submissions will not be accepted).
- All assignments must be submitted via CourseWeb.

Assignments

- If submitting multiple files, they must be zipped into a single file using standard .ZIP format. The final zipped file must be titled with the last names of the author, number of the assignment and course number separated by underscores. For example, if your last name is Smith, and you are submitting assignment 2, your final file should be named **Smith_Assignment2_INFSCI2710.zip**.
- You will lose 2 points for every submission that does not follow this naming convention.

Assignments

- Four Homework
- One group project

Late Submissions

Projects/assignments submitted after due date will be accepted, but your overall grade for that project/assignment will be reduced by 25% of the grade after the submission deadline.

In-Class Activities

- Some in-class activities
- An Educational Data-Driven Course
 - SQL Practice System
 - Play/Practice some examples in the system (from week 4 to 7).
 - Weekly participation = 0.5 point (total 2 points)
 - More information will be provided later.

Grading Scale

Grade	Range of Scores	Grade	Range of Scores
A+	95-100	C+	65-69
A	90-94	C	60-64
A-	85-89	C-	50-60
B+	80-84	D	20-50
B	75-79	F	<20
B-	70-74		

Collaboration vs. Cheating

Collaboration on homework is permitted to an extent. Specifically, students are allowed to discuss the possible solutions to a problem and help each other with logic errors. However, handing your work to someone so that they may see a copy of your solution, or dictating code to a person on line-by-line basis is not within the spirit of the collaboration policy or the honor code of the university.

Academic Integrity Statement

Cheating/plagiarism will not be tolerated. All work must be your own, unless collaboration is specifically and explicitly permitted as in the course group project. Any unauthorized collaboration or copying will at minimum result in no credit for the affected assignment and may be subject to further action under the University Guidelines for Academic Integrity (<http://www.provost.pitt.edu/info/ai1.html>). You may incorporate excerpts from publications by other authors, but they must be clearly marked as quotations and properly attributed. You may discuss your ideas with others, but all substantive writing and ideas must be your own, or else be explicitly attributed to another, using a citation sufficiently detailed for someone else to easily locate your source.

Disability

If you have a disability for which you are or may be requesting an accommodation, you are encouraged to contact the Instructor and Disability Resources and Services, 216 William Pitt Union, (412) 648-7890 / (412) 383-7355 (TTY), as early as possible in the term. Disability Resources and Services reviews documentation related to a student's disability, provides verification of the disability, and recommends reasonable accommodations for specific courses.

Introduction to course and syllabus overview

What is Data and Why Do We Care?

Database Management Systems overview

Tools overview

What is Data?

What is data?

1. factual information (such as measurements or statistics) used as a basis for reasoning, discussion, or calculation
2. information output by a sensing device or organ that includes both useful and irrelevant or redundant information and must be processed to be meaningful
3. information in numerical form that can be digitally transmitted or processed

<https://www.merriam-webster.com/dictionary/data>

What is data?

- Data is a set of values of **qualitative** or **quantitative** variables.
- Pieces of data are individual pieces of information.
- While the concept of data is commonly associated with scientific research, data is collected by a huge range of organizations and institutions, including businesses, governments, etc...

<https://en.wikipedia.org/wiki/Data>



What is Data Mining?

What is data mining?

The use of **machine learning algorithms** to find **patterns** of relationship between data elements in large, noisy and messy datasets, which can lead to **knowledge discovery**

What is Knowledge Discovery?

What is knowledge discovery?

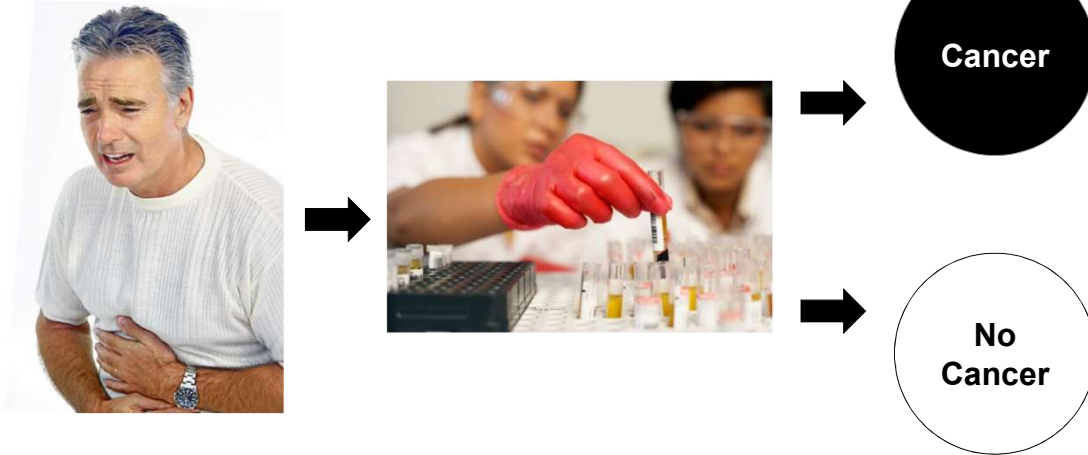
The broad process of understanding **patterns** and their **meaning** in data

What is Machine Learning?

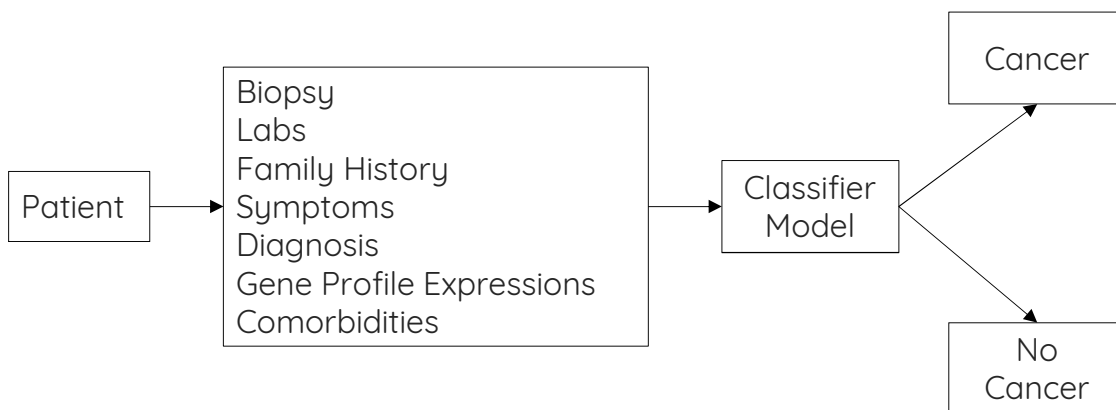
What is machine learning?

- Machine learning is a method of data analysis that automates analytical model building.
- Using algorithms that iteratively learn from data, machine learning allows computers to find hidden insights without being explicitly programmed where to look.

The Big Question



The Big Question



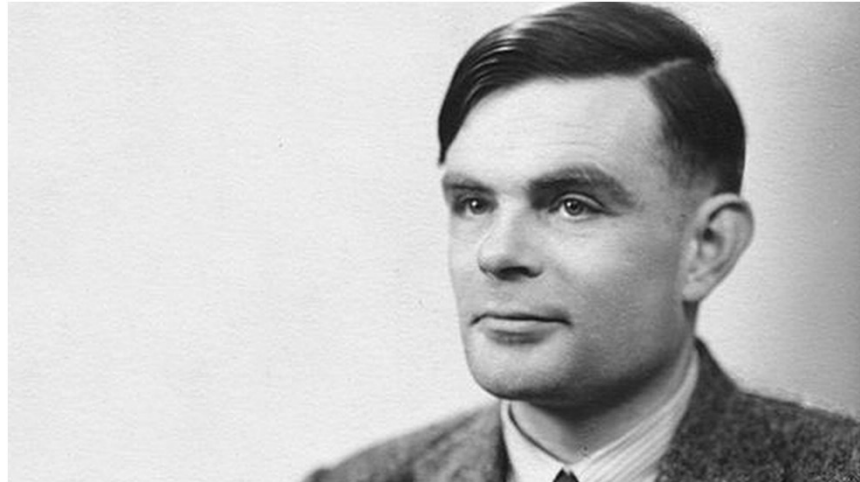
What is Artificial Intelligence?

What is artificial intelligence?

Artificial Intelligence (AI) is usually defined as the science of making computers do things that require intelligence when done by humans.

http://www.alanturing.net/turing_archive/pages/reference%20articles/what%20is%20ai.html

Turing Test



Demo

Luna: Open Source Artificial Intelligence Demo:
<https://www.youtube.com/watch?v=GeSqLMiKhBA>

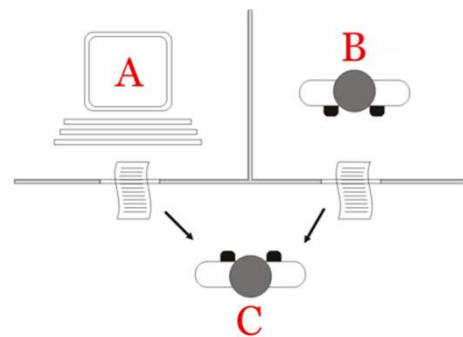
Turing Test

- Developed by Alan Turing in 1950
- Test of a machine's ability to exhibit intelligent behaviour equivalent to, or indistinguishable from, that of a human.

https://en.wikipedia.org/wiki/Turing_test

Turing Test

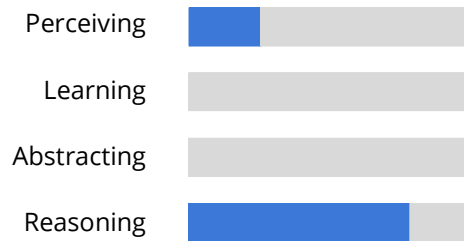
- Player C, the interrogator, is given the task of trying to determine which player - A or B - is a computer and which is a human.
- The interrogator is limited to using the responses to written questions to make the determination.



https://en.wikipedia.org/wiki/Turing_test

First Wave of AI

- Handcrafted Knowledge
- Rule-Based Systems



<https://www.youtube.com/watch?v=-O01G3tSYpU>

Chess

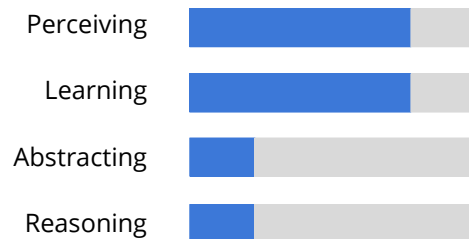


Failures



Second Wave of AI

- Statistical modeling
- Machine learning
- Big Data
- Examples: voice recognition, face recognition, image classification



Second Wave of AI



← **CAT**

Second Wave of AI

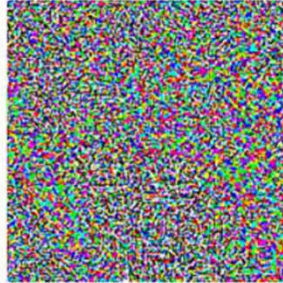


← **?**

Fails



+ .007 ×



=



“panda”
57.7% confidence

“nematode”
8.2% confidence

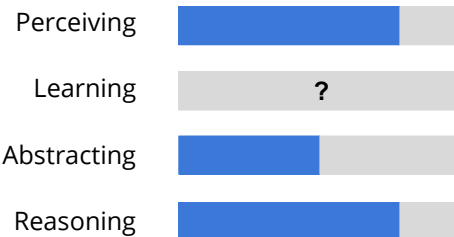
“gibbon”
99.3 % confidence

<http://www.popsci.com/byzantine-science-deceiving-artificial-intelligence>

Third Wave

→ Contextual Adaptation

- ◆ Systems construct explanatory models for classes of real world phenomena
- ◆ Understand why / why not
- ◆ Contextual models



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

Third Wave of AI



← I know it's
a cat - now
tell me
WHY!

Let's talk about
dangers of data



Correlations

- Correlation does not mean causation!
- Not all correlations are meaningful
- Some correlations are anomalous or spurious
- Too many correlations

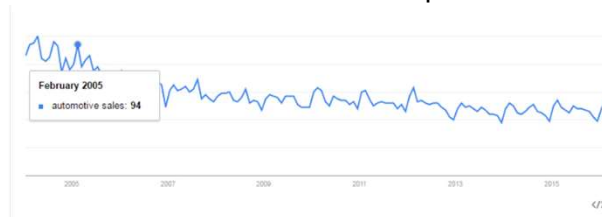
Gaming The System

- Gaming automated essay graders
- Search engine optimization (SEO)
- Artificially enhancing citations
- Leveraging social media to “trend” specific subjects
- Examples of ‘J.C. Penney scandal’*

*https://www.huffingtonpost.com/2014/03/19/jcpenney-prices_n_4986649.html

Interpreting Results

- Often even precise answers do not enhance our knowledge.
- Google Trends search for "Automotive Sales" produces a graph that shows a somewhat cyclical nature of interest over time in automotive sales.
- We should not jump to conclusions about what causes this particular trend and its implications.



Data Mining Process

- 1. Data preprocessing**
- 2. Explore your data**
3. Generate summary statistics, basic visualization
4. Select appropriate data points (variables)
5. Deal with problem values (outliers and missing values)
6. Define data mining methods appropriate for your data and your research question
7. Partition data into training and evaluation subsets
8. Confirm (validate) your findings
9. Suggest actions from the data mining results

Introduction to course and syllabus overview

What is Data and Why Do We Care?

Database Management Systems overview

Tools overview

What is a server?

What is the difference between a spreadsheet and a database?

Databases are safer. Excel, for example, does everything in memory, so that any unsaved data may be lost if your system crashes. Databases write data to the hard drive immediately.

<http://www.pcmag.com/article2/0,2817,1435148,00.asp>

Databases can handle more data. Sure, Excel can technically handle more than 65,000 rows of data, but doing so will likely bog down even the fastest PC.

<http://www.pcmag.com/article2/0,2817,1435148,00.asp>

Databases can easily link tables of related data together, such as customers and orders or musical groups and albums (as well as the songs on each album). This is where the words *relational* and *database* come together. Storing related data together in a single table or spreadsheet can be unwieldy and invite errors.

<http://www.pcmag.com/article2/0,2817,1435148,00.asp>

Most importantly, databases can be used to answer complex questions.

Patient	Date	Symptom	Country
1	07/12/2014	cough, fever	Guinea
2	07/12/2014	cough with blood production, diarrhea, fever	Liberia
3	07/13/2014	reddened eyes, joint and muscle pain, fever	Liberia
4	07/13/2014	fever, fatigue, weakness, reddened eyes	Liberia
5	07/13/2014	joint and muscle pain, headache, nausea and vomiting	Sierra Leone
6	07/13/2014	fever, fatigue, malaise, and weakness, reddened eyes, joint and muscle pain, headache, nausea and vomiting	Guinea
7	07/13/2014	fever, fatigue, weakness, reddened eyes	Sierra Leone
8	07/14/2014	joint and muscle pain, headache, nausea and vomiting	Guinea
9	07/14/2014	cough with blood production, diarrhea, fever	Liberia
10	07/14/2014	fever, fatigue, malaise, and weakness, reddened eyes, joint and muscle pain, headache, nausea and vomiting	Liberia

Types/Brands of Relational DBMS

- Microsoft SQL Server (a.k.a MSSQL)
- Oracle
- MySQL (*our focus in this class*)
- Microsoft Access
- PostgreSQL

Microsoft SQL Server



<http://www.microsoft.com/en-us/server-cloud/products/sql-server-editions/sql-server-express.aspx>

Oracle



<http://www.oracle.com/technetwork/database/enterprise-edition/downloads/index.html>

MySql



<http://www.mysql.com/>

Microsoft Access



<http://office.microsoft.com/en-us/access/>

PostgreSQL



<http://www.postgresql.org/>

Database Management System (DBMS)

- DBMS contains information about a particular **enterprise**
 - Collection of interrelated data
 - Set of programs to access the data
 - An environment that is both *convenient* and *efficient* to use

Database Applications

- Banking: all transactions
- Airlines: reservations, schedules
- Universities: registration, grades
- Sales: customers, products, purchases
- Online retailers: order tracking, customized recommendations
- Manufacturing: production, inventory, orders, supply chain
- Human resources: employee records, salaries, tax deductions

Entities

- A concept in the business or user environment about which the organization wishes to maintain data
- Person, place object, event, or concept

Entities Examples

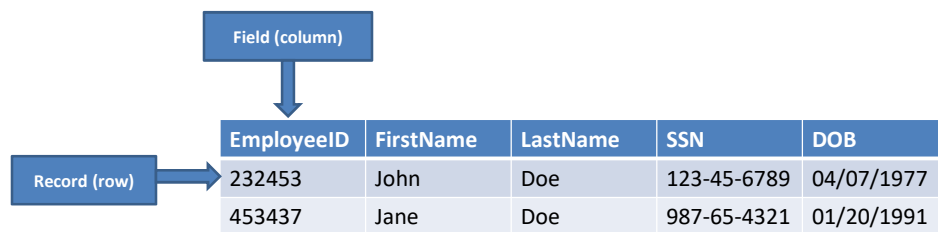
- Person: Employee, Student, Patient
- Place: Store, Warehouse, State
- Object: Machine, Building, Automobile, Product
- Event: Sale, Registration
- Concept: Account, Course

Attributes

- An attribute is a property or characteristic of an entity
- Example: Model, make, year, color are attributes of **Car** entity.

Tables

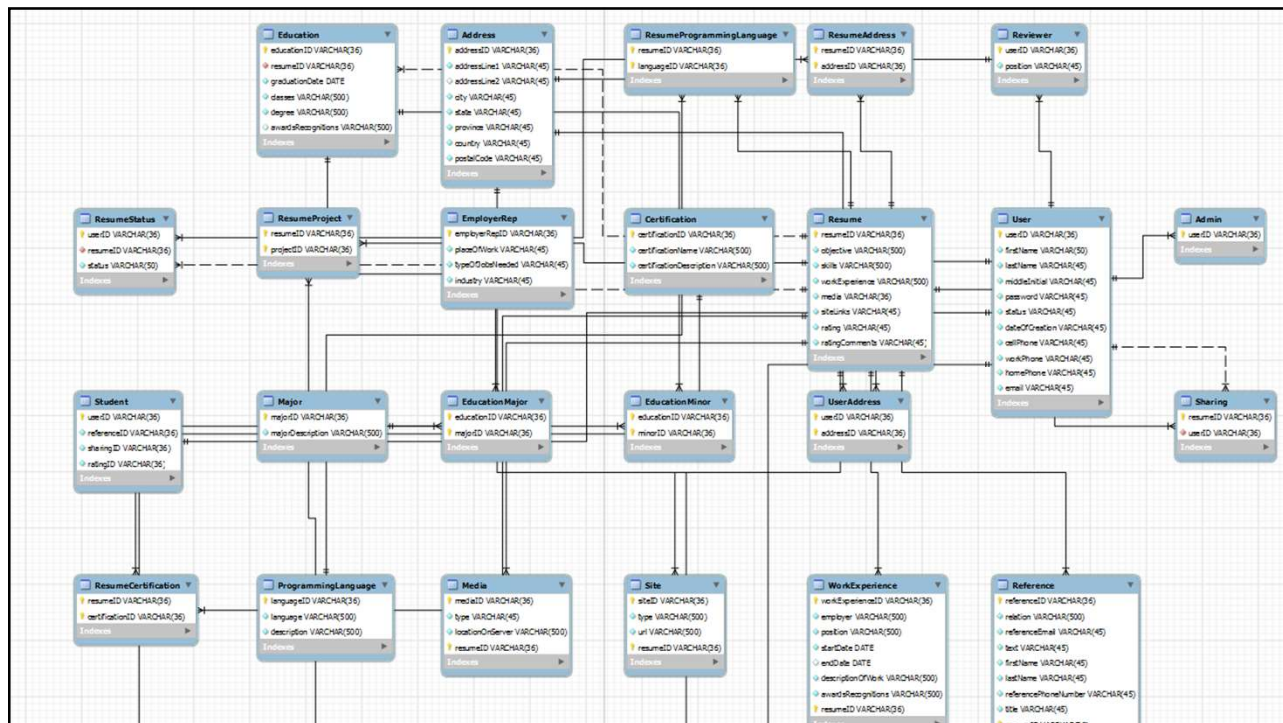
- A database is a collection of tables.
- Tables are referred to as “**entities**”.
- Each table contains records (or tuples) - horizontal rows in the table.
- Each record contains fields - vertical columns of the table.
- Columns are referred to as “**attributes**”



EmployeeID	FirstName	LastName	SSN	DOB
232453	John	Doe	123-45-6789	04/07/1977
453437	Jane	Doe	987-65-4321	01/20/1991

Schema

- A collection of related tables and relationships between those tables is called a **schema**.
- Some database management applications (such as Microsoft SQL Server) allow multiple schemas per database



SQL

- SQL stands for Structured Query Language
- **Semi-standardized** language for querying relational databases
- SQL differs slightly between database brands
- **SQL != Database**

Query to select first 5 rows from a database table called Employees:

- Microsoft SQL Server:
 - *SELECT **TOP 5** * FROM Employees*
- MySQL
 - *SELECT * FROM Employees **LIMIT 5**;*

Database Design

- The process of designing the general structure of the database consists of:
 - Logical design
 - Physical design

Logical Design

- Logical Design – deciding on the database schema.
Database design requires that we find a “good” collection of relation schemas.
 - **Business decision** – What attributes should we record in the database?
 - **Computer Science decision** – What relation schemas should we have and how should the attributes be distributed among the various relation schemas?

Physical Design

- Physical Design – Deciding on the physical layout of the database
 - File system
 - Indexes

Database Architecture

The architecture of a database systems is greatly influenced by the underlying computer system on which the database is running:

- Centralized
 - Client-server (*our focus in this class*)
 - Parallel (multi-processor)
 - Distributed
-
- This class will not cover the topic of data warehouse and nosql.

Data Warehouses

- System used for reporting and data analysis
- Considered a core component of business intelligence
- Central repositories of integrated data from one or more disparate sources.
- Store current and historical data in one single place
- Used for creating analytical reports for knowledge workers throughout the enterprise

https://en.wikipedia.org/wiki/Data_warehouse

Data Warehouses

- The data stored in the warehouse is uploaded from the operational systems (such as marketing or sales).
- The data may pass through an operational data store and may require data cleansing for additional operations to ensure data quality before it is used in the DW for reporting.

https://en.wikipedia.org/wiki/Data_warehouse

NoSQL?

NoSQL

- **Document databases** pair each key with a complex data structure known as a document. Documents can contain many different key-value pairs, or key-array pairs, or even nested documents.
- **Graph stores** are used to store information about networks of data, such as social connections. Graph stores include Neo4J and Giraph.

NoSQL Data

- Great for unstructured data:
 - Emails
 - Text files
 - Spreadsheets
 - Digital Images
 - Video
 - Audio
 - Social media posts
- List of NoSQL databases: <http://nosql-database.org/>

The Benefits of NoSQL

- More scalable than relational DBs
- Provide superior performance
- Their data model addresses large volumes of rapidly changing structured, semi-structured, and unstructured data
- Object-oriented programming that is easy to use and flexible
- Geographically distributed scale-out architecture instead of expensive, monolithic architecture

<https://www.mongodb.com/nosql-explained>

Document Databases

Designed for storing, retrieving, and managing document-oriented, or semi structured data.

```
{
  _id: <ObjectId>,
  username: "123xyz",
  contact: {
    phone: "123-456-7890",
    email: "xyz@example.com"
  },
  access: {
    level: 5,
    group: "dev"
  }
}
```

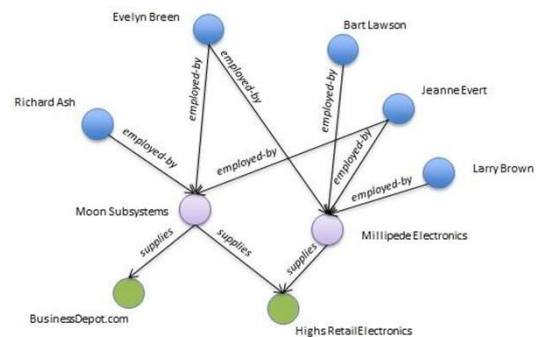
Embedded sub-document

Embedded sub-document

<https://docs.mongodb.com/manual/core/data-modeling-introduction/>

Graph Databases

- Use graph structures for semantic queries with nodes, edges and properties to represent and store data.
- A key concept of the system is the graph where entities are **nodes** and relationships are **edges**.



<https://upside.tdwi.org/articles/2016/07/25/graph-databases-for-analytics-1.aspx>

Introduction to course and syllabus overview

What is Data and Why Do We Care?

Database Management Systems overview

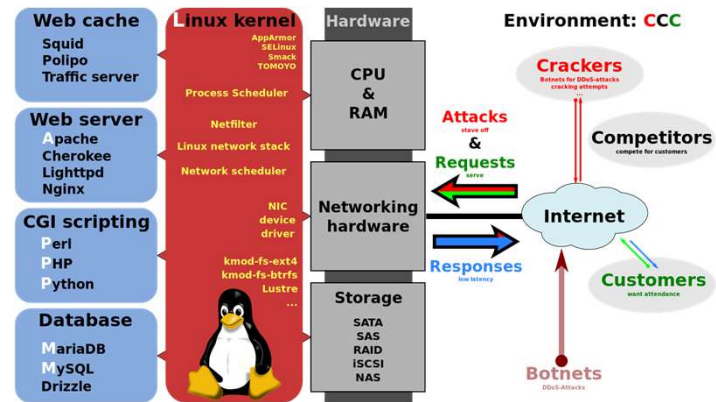
Tools overview

Tools

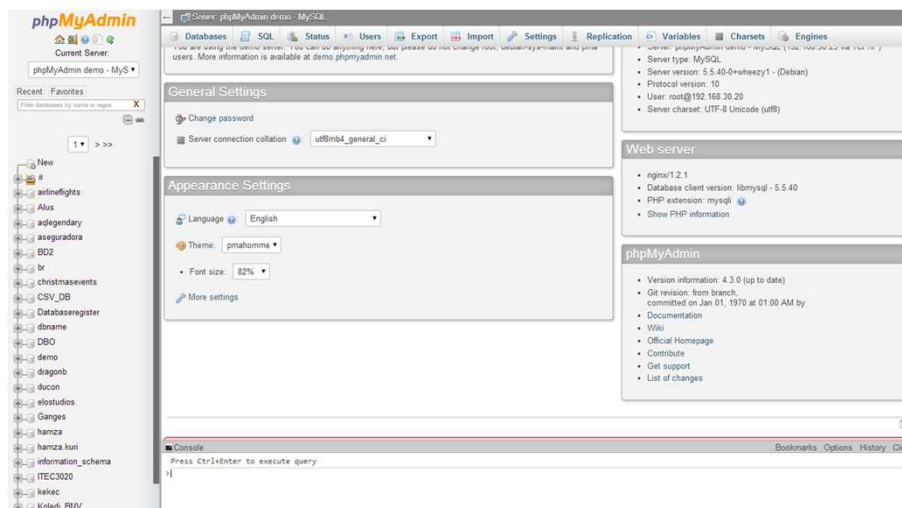
- Bring your laptop/tablet.
- Many in-class practices.

Installing LAMP

- WAMP for Windows
- MAMP for Mac
- LAMP for Linux.



phpMyAdmin



Draw.IO

- Login to Google Drive (<https://drive.google.com>)
- Go to New → More → Connect More Apps
- Find draw.IO app
- Connect it to your Google Drive

Resources

- [SQL Tutorial - W3Schools](#)
- [PHP 5 Tutorial - W3Schools](#)

Questions?