DS501: Midterm Review!

Prof. Randy Paffenroth
rcpaffenroth@wpi.edu

Worcester Polytechnic Institute

WPI

Rules for midterm exam

The midterm exam and final exam will be in class, noncumulative, and open note, but **no collaboration will be allowed** and the exams be graded based upon demonstrated understanding of key concepts. For each exam, you are allowed to bring in up to **4 (four)** 8 ½ by 11 sheets of paper (either printed or handwritten) with whatever notes you want for the exam. You are also allowed a calculator that does not have a network connection (e.g., **no cell phones!**).

WPI

Class 1
Introduction

WPI

Ok, let the good time roll!
Oh, wait...

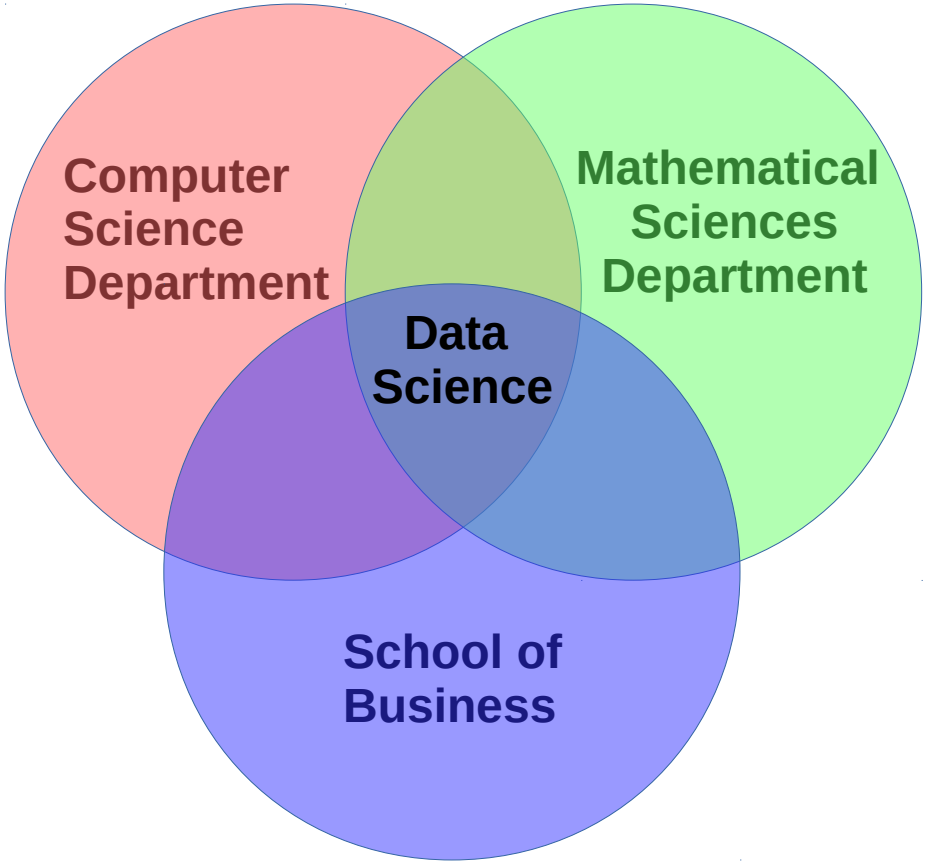
*Given a large mass of data, we can by judicious selection construct **perfectly plausible** unassailable theories—*
***all** of which, **some** of which, or **none** of which may be right.*
- Paul Arnold Srere

*"It's tough to make **predictions**, especially about the future."*

— Yogi Berra

WPI

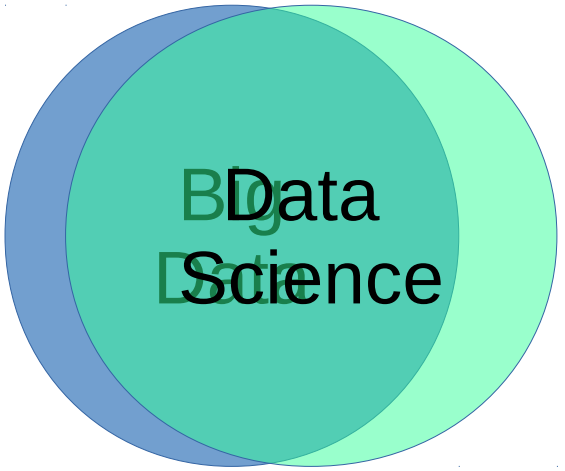
What is Data Science? Another view...



- Based upon Drew Conway's Data Science Venn Diagram
  - http://en.wikipedia.org/wiki/Data_science
  - http://drewconway.com/zia/2013/3/26/the-data-science-venn-diagram

What do they have to do with each other?

- Are Big Data and Data Science the same thing?
  - I wouldn't say so...
  - Data Science can be done on small data sets.
  - And not everything done using Big Data would necessarily be called Data Science.
  - But there certainly is a substantial overlap!



So, there is one thing that I really wany to stress.

- The three V's are all extremely important and make a Data Scientists job interesting and important, but I want to push for a 4th V!

  - Veracity:
    - Ok you have made a prediction, do you bet the farm (or your job on it)?
    - Or, maybe you do an *experiment* to see if the predictions you are making are correct?
    - Is one experiment enough to bet the farm?
    - Is a million?
    - How do you ***think critically*** about data, and all the things that go along with it?

What is Big Data?

- The are many examples of "data", but what makes some of it “big”? The classic definition revolves around the **three Vs**.
- **Volume, velocity, and variety.**
  - **Volume:** There is a just a lot of it being generated all the time. Things get interesting and “big”, when you can’t fit it all on one computer anymore. Why? There are many ideas here such as MapReduce, Hadoop, etc. that all revolve around being able to process data that goes from Terabytes, to Petabytes, to Exabytes.
  - **Velocity:** Data is being generated very quickly. Can you even store it all? If not, then what do you get rid of and what do you keep?
  - **Variety:** The data types you mention all take different shapes. What does it mean to store them so that you can play with or compare them?

http://pl.wikipedia.org/wiki/Green_Giant#mediaviewer/Plik:Jolly_green_giant.jpg

Class 2
Data Gathering

Methods of data collection

- Surveys
  - Asking people what they think
- Experiments
  - Making your own little world and seeing what you can find our
- Observation
  - Looking at the world that is and seeing that is there

Types of Biases

Bias can occur in a sample due to various reasons as follows :

1. **Sampling Bias**: As the term suggests, this kind of bias results from a flaw in the sampling method, most likely if the sample is **non-random**. Another way it can occur is due to **under-coverage** – having a sample that lacks representation from parts of the population. Responses by those not in the sample might be quite different from those in it, thus leading to misleading conclusions about the population.

Example: A telephone survey will not reach homeless people; incidentally, these groups of people may have very different views about life in general.

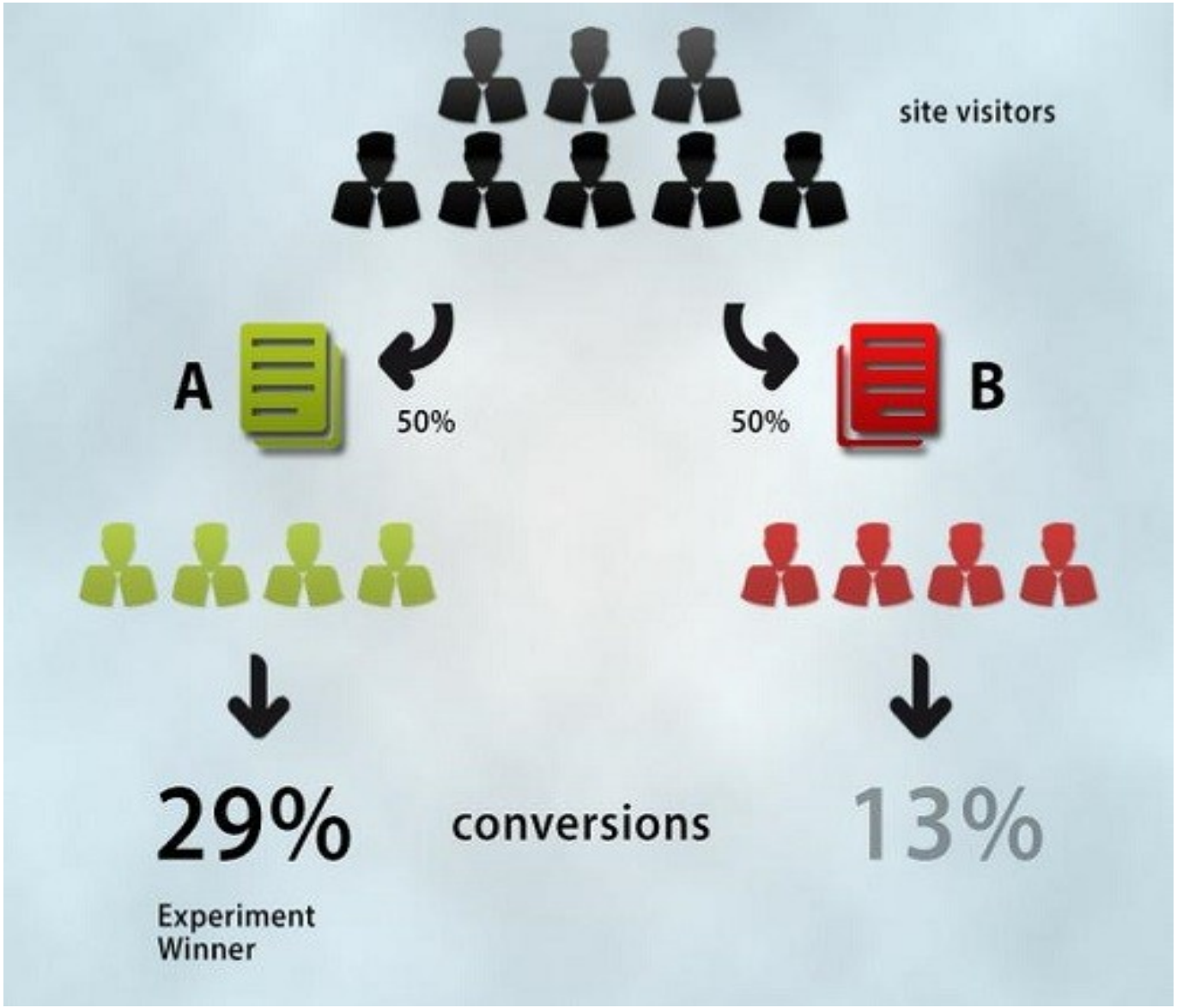
Types of Biases

2. **Non-response bias**: This kind of bias results when some of the sampled subjects cannot be reached or refuse to participate. In fact, the subjects who are willing to participate may be different from the overall sample in some way, perhaps having strong views about the survey issues. The subjects who do participate may not respond to some questions, resulting in non-response bias due to missing data.

Types of Biases

3. **Response bias**: This kind of bias results from the actual responses. The responses of subjects may differ based on the particular manner ***the interviewer asks questions;*** subjects can often lie because they think that their responses may be socially unacceptable.

Inspirational example for experimental study: A/B testing



https://www.flickr.com/photos/rahulrodriguez/14683524180

What About Data Quality?

- Generally, you have a problem if the data doesn't mean what you think it does, or should
  - Data not up to spec : garbage in, glitches, etc.
  - You don't understand the spec : complexity, lack of metadata.
- Data quality problems are expensive and pervasive
  - DQ problems cost hundreds of billion $$$ each year.
  - Resolving data quality problems is often the biggest effort in a data mining study.

Adapted from: "Data Quality and Data Cleaning: An Overview"
Theodore Johnson, SIGMOD 2003

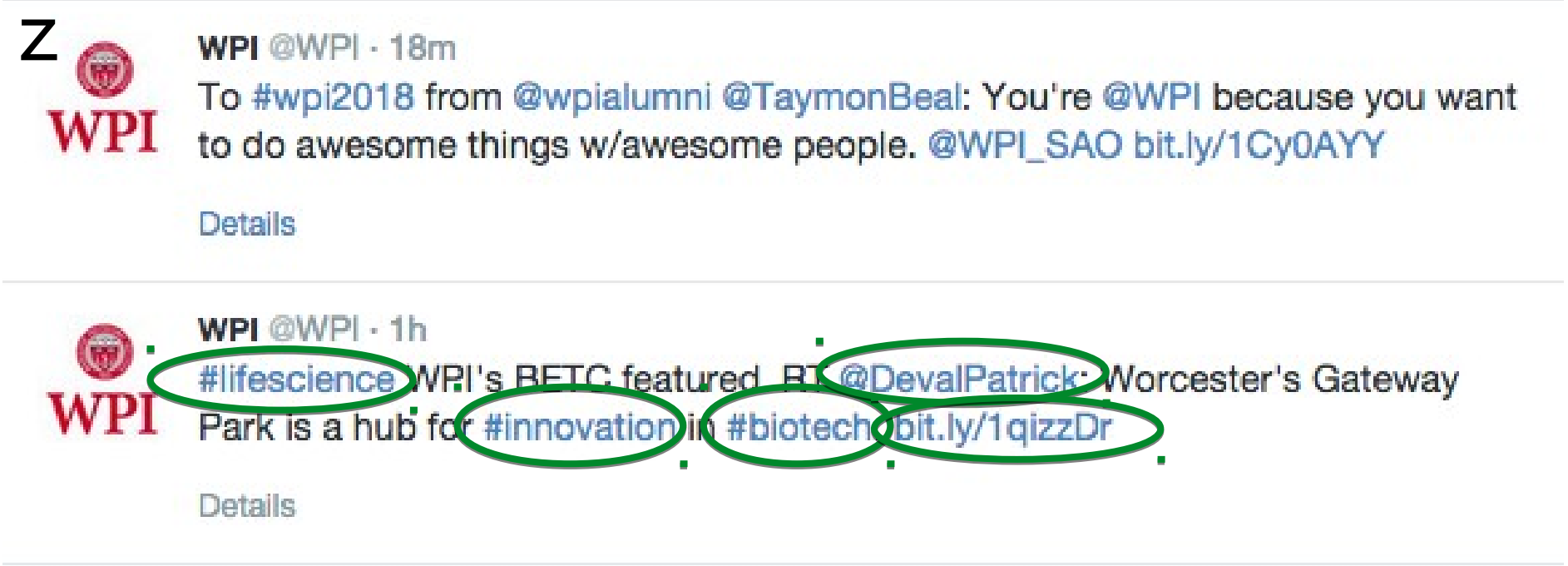
One way to get data: Web page "crawling"

- HTML is all about how to display/show data, but not about giving you the data.
- Easy to download, but, hard to process
- Powerful, but can actually be quite complicated to get data from web sites.
- Rules: robots.txt

Learn about the Data

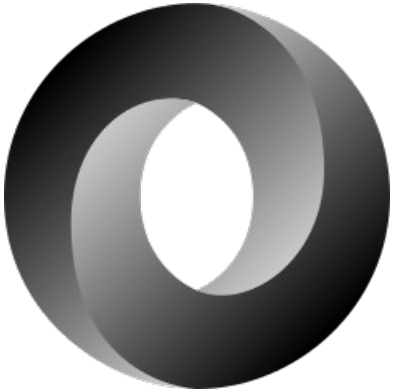
https://support.twitter.com/articles/166337?lang=en#

- **Twitter Data**
  - Tweets: 140 characters (text + entities)



https://twitter.com/search?q=wpi

JSON

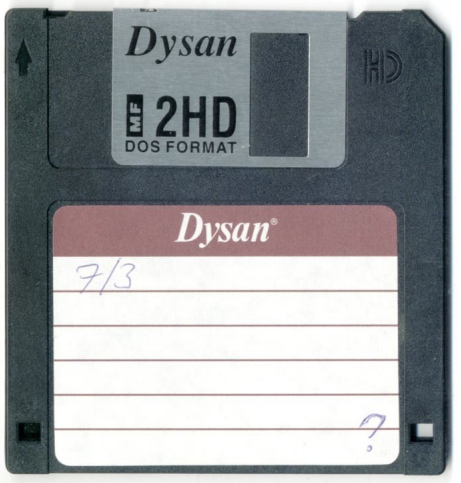


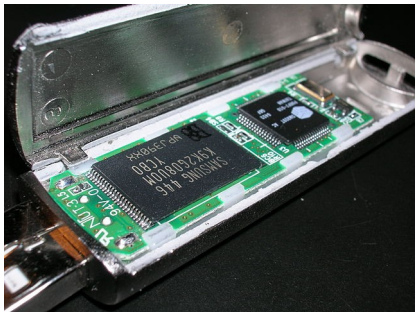
- **JavaScript Object Notation (JSON)**
- An open standard format that uses human-readable text to transmit data objects consisting of attribute–value pairs.
- A list of Dictionaries

```
{
  "firstName": "John",
  "lastName": "Smith",
  "age": 25,
  "address": {
    "streetAddress": "21 2nd Street",
    "city": "New York",
    "state": "NY",
    "postalCode": "10021"
  },
  "phoneNumber": [
    {
      "type": "home",
      "number": "212 555-1239"
    },
    {
      "type": "fax",
      "number": "646 555-4567"
    }
  ],
  "gender": {
    "type": "male"
  }
}
```

Class 3
Data Storage

Where do we store data?





"USB flash drive". Licensed under CC BY-SA 3.0 via Commons - https://commons.wikimedia.org/wiki/File:USB_flash_drive.JPG#/media/File:USB_flash_drive.JPG

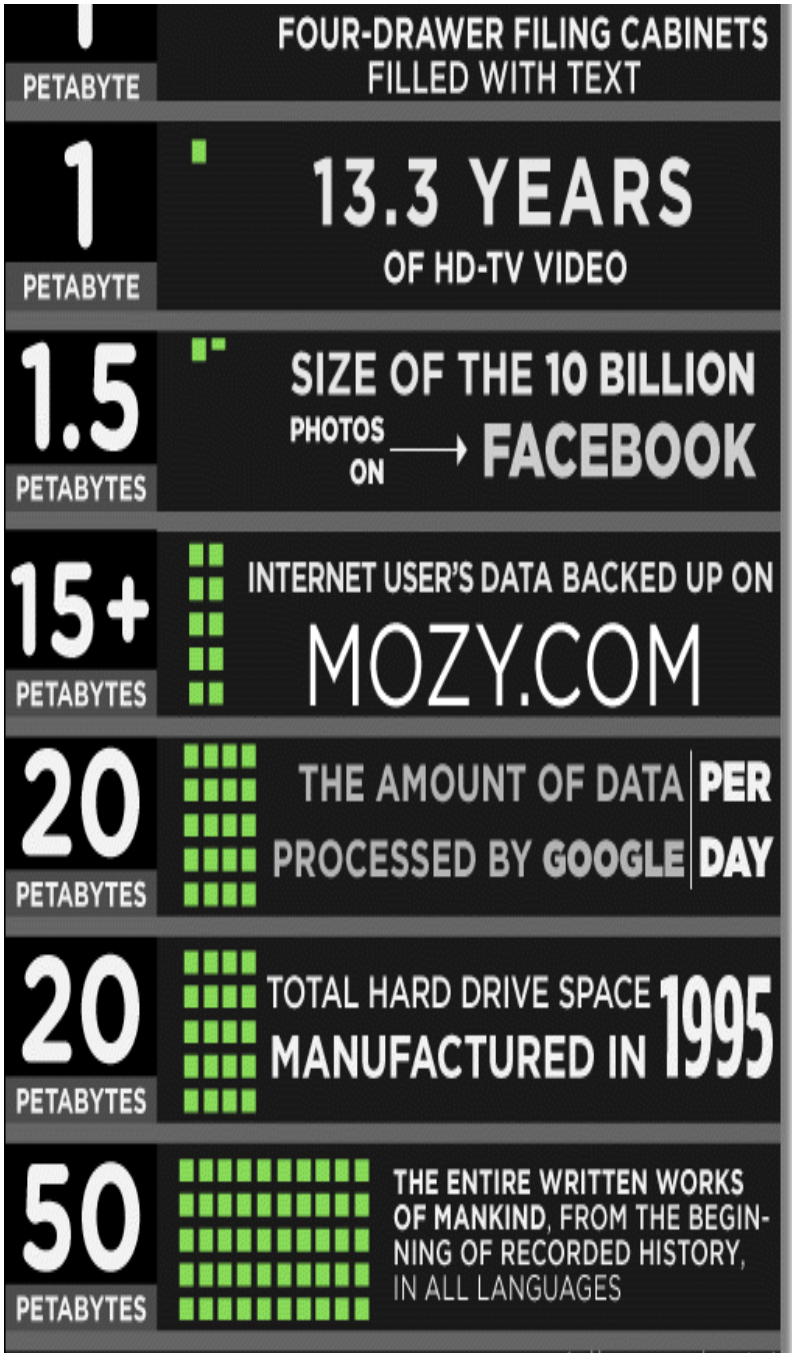
"Laptop-hard-drive-exposed" by Evan-Amos - Own work. Licensed under CC BY-SA 3.0 via Commons - https://commons.wikimedia.org/wiki/File:Laptop-hard-drive-exposed.jpg#/media/File:Laptop-hard-drive-exposed.jpg

But, how can we all participate? Example...

Amazon AWS S3

"AWS Simple Icons Storage Amazon S3 Bucket with Objects" by Amazon Web Services LLC - http://aws.typepad.com/aws/2011/12/introducing-aws-simple-icons-for-your-architecture-diagrams.html. Licensed under CC BY-SA 3.0 via Wikimedia Commons - https://commons.wikimedia.org/wiki/File:AWS_Simple_Icons_Storage_Amazon_S3_Bucket_with_Objects.svg#/media/File:AWS_Simple_Icons_Storage_Amazon_S3_Bucket_with_Objects.svg

https://aws.amazon.com/s3/pricing/



1 PetaByte

http://mswhs.files.wordpress.com/2009/07/whatsapetabyte.gif

What is a Database System?

- A **database** is an organized **collection** of data.
  - The focus: efficient data query/ retrieval.
- A "database management system" (DBMS) is a suite of computer software providing the interface between users and a database or databases.

source from: wikipedia

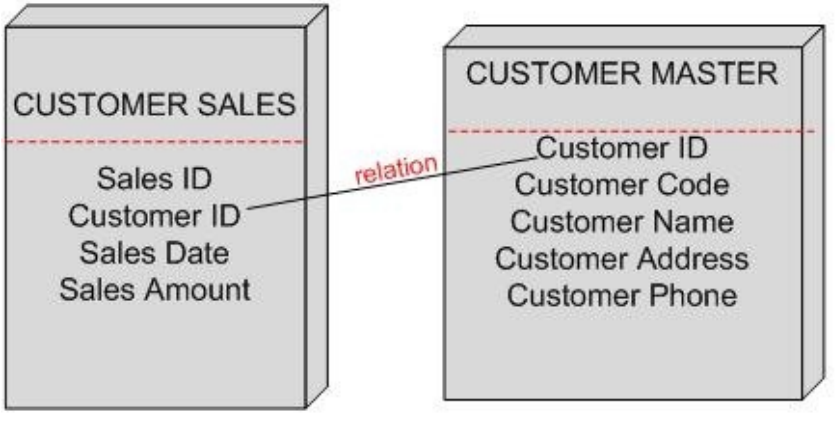
Problems DBMS can solve

- **Data Model**: clean and organized data
- **Scale**: too large to fit in memory
- **Sharing**: multiple readers and writers
- Concurrent access, recovery from crashes
- Reduced application development time

Relational Model

Main concept: **relation**, basically a table with rows and columns.

Every relation has a **schema**, which describes the columns, or fields



Sales ID	Customer ID	Sales Date	Sales Amount
1	101	12/09/2008	10000
2	101	01/09/2008	23789
3	102	02/07/2008	45000
4	103	11/06/2008	25345

Customer ID	Customer Code	Customer Name	Customer Address	Customer Phone
101	C00101	All sec Corp	Houston, Texas	001-325-789-321
102	C00102	John S	Chennai	0091-44-273910
103	C00103	Bridge Inc.	Delhi	0091-11-456801
104	C00104	Symphony Org	Bombay	0091-22-568902

JSON

- Data model for semi-structured data

```
{ "users":[
            {
                "firstName":"Ray",
                "lastName":"Villalobos",
                "joined": {
                    "month":"January",
                    "day":12,
                    "year":2012
                }
            },
            {
                "firstName":"John",
                "lastName":"Jones",
                "joined": {
                    "month":"April",
                    "day":28,
                    "year":2010
                }
            }
    ]}
```

- Basic types: number, string, boolean, …
- Objects { }
  - sets of label-value pairs
- Arrays [ ]
  - list of values

Practice

Which is NOT a valid JSON object?

```
{ "name":
"Smiley",
  "age": 20,
  "phone": null,
  "email": null,
  "happy": true
}
```

```
{ "name": "Smiley",
  "age": 20,
  "phone": "888-123-
4567",
  "email":
"smiley@xyz.com",
  "happy": true }
```

```
{ "name": "Smiley",
  "age": 20,
  "phone": "888-123-
4567",
  "email":
smiley@xyz.com,
  "happy": true }
```

```
{ "name":
"Smiley",
  "age": 20,
  "phone": null,
  "email":
"null",
  "happy": true
}
```

Practice

Which is NOT a valid JSON array?

```
[ [1, 2], ["dog", "cat"], [true, false], [1, "dog",
null],
  {"pet":"dog", "fun":true} ]
```

```
[ 1, 2, "dog", "cat", true, false,
[],
  {"pet":"dog", "fun":true} ]
```

```
[ 1, 2, dog, cat, true, false, [1, "dog",
null],
  {"pet":"dog", "fun":true} ]
```

```
[ 1, 2, "dog", "cat", true, false, [1, "dog", null],
{} ]
```

XML

- Extensible Markup Language (XML)

```
<?xml version="1.0" standalone="yes"?>
<BankAccount>
      <Number>1234</Number>
      <Type>Checking</Type>
      <OpenDate>11/04/1974</OpenDate>
      <Balance>25382.20</Balance>
      <AccountHolder>
            <LastName>Singh</LastName>
            <FirstName>Darshan</FirstName>
      </AccountHolder>
</BankAccount>
```

- HTML (format), XML (content)

Comparison of Data Models

	Relational	JSON	XML
Structure	Structured	Semi-structured	Hierarchical, Tree
Schema	Fixed	flexible	flexible, “self-describing”
Query	simple, easy	not as easy	less so
Ordering	BasedSet	Arrays	Implied ordering
Implementation	Native Systems	NoSQL systems	Various

Goals of Database systems

- All users to ***create*** new databases and specify their schema (using a data-definition language)

- Give users the ability to ***query*** and modify the data (using a query language or data-manipulations language)

- Support storage of very large amounts of data (terabytes or more)

- Enable ***durability***, recover after failures, errors

- Control access to data from many users. Each user should work in *isolation*.

Transaction Processing

Group of one or more database operations into a transaction which must be executed atomically and in isolation of other trans.

- ACID properties of Transactions
  - A: “atomicity”, all-or-nothing execution of transactions
  - C: “consistency”, constraints on data, transactions are expected to preserve consistency.
  - I: “isolation”, no other transactions appears to be executed at the same time
  - D: “durability”, the effect of a transaction must never be lost, once the transaction has completed

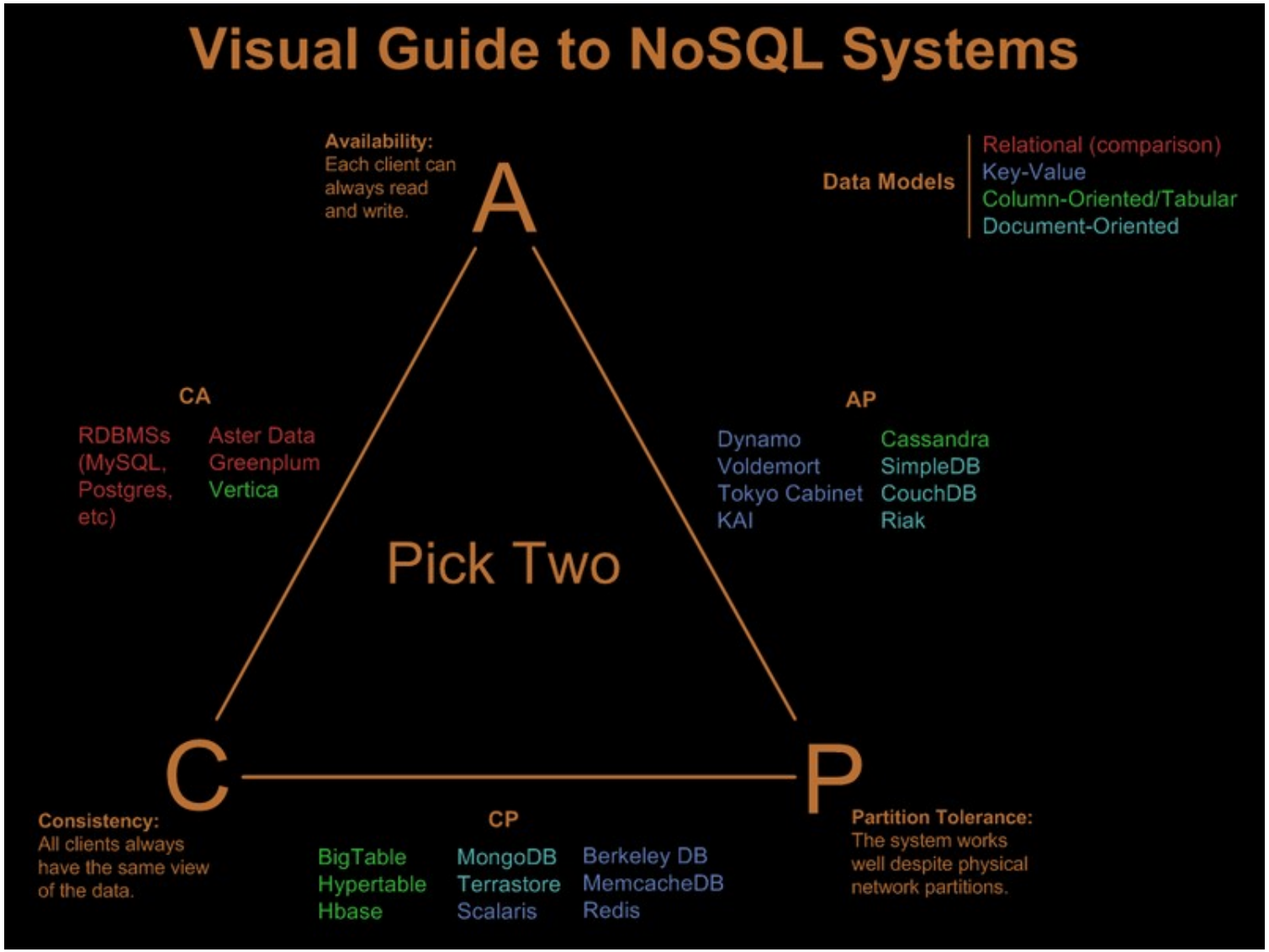
Today NoSQL!

- "Non-SQL"
- "Non-relational"
- "Not only SQL"

Trade-offs

- Cap Theorem or Brewer's Theorem (Eric Brewer)
- It is impossible for a distributed computer systems to simultaneously provide all three of the following guarantees:
  - **Consistency** (all nodes see the same data at the same time)
  - **Availability** (a guarantee that every request receives a response about whether it succeeded or failed)
  - **Partition tolerance** (the system continues to operate despite arbitrary partitioning due to network failures)

NoSQL systems



http://blog.nahurst.com/visual-guide-to-nosql-systems

Class 4
Business Intelligence

Business Intelligence

- What did we talk about?
  - Definition
  - Focused on three pieces
  - Three case studies

WPI

Definition

Business Intelligence is a user-oriented process of gathering, exploring, interpreting and analyzing of data, which leads to the streamlining and rationalization of the *decision-making* process. Those systems support managers in business *decision-making* in order to create economy value growth of an enterprise.

- Business Intelligence: Making Decisions through Data Analytics

WPI

Our three main topics for each case study

- To keep things simple, we will focus on three important parts of business intelligence.
  - Measurements and data gathering
  - Data analysis and exploration
  - Distribution, reporting and data visualization.
- In particular, we showed how different types of businesses go through these stages and make business decisions.

WPI

Three case studies

- "Appeltaart" by Original uploader was BlueBart at nl.wikipedia - Transfered from nl.wikipedia. Licensed under Creative Commons Attribution 1.0 via Wikimedia Commons - http://commons.wikimedia.org/wiki/File:Appeltaart.jpg#mediaviewer/File:Appeltaart.jpg

"Netflix logo" by Netflix - Netflix Media CenterTransferred from en.wikipedia to Commons by User:SethAllen623 using CommonsHelper.. Licensed under Public domain via Wikimedia Commons - http://commons.wikimedia.org/wiki/File:Netflix_logo.svg#mediaviewer/File:Netflix_logo.svg

"International Wal-Mart Truck" by Amanda Bengtson - Flickr: International Wal-Mart Truck. Licensed under Creative Commons Attribution-Share Alike 2.0 via Wikimedia Commons - http://commons.wikimedia.org/wiki/File:International_Wal-Mart_Truck.jpg#mediaviewer/File:International_Wal-Mart_Truck.jpg

WPI

Things to think about

- Can you define **Business Intelligence**?
- Given a company, can you you:
  - List data they might **gather**?
  - Think about how they would **analyze** the data they gather?
  - Describe how the data can be **distributed and visualized** in the company?
- Given two companies:
  - Can you reason about which of the three parts of business intelligence is **easier** for one company than it is for the other?
  - Can you reason about which of the three parts of business intelligence is **harder** for one company than it is for the other?

Class 5
Basic Statistics and Probability

Basic Statistics and Probability

- What did we talk about?
  - Discrete random variables
  - Continuous random variables
  - Conditional probabilities
  - Bayes Theorem
  - The Base Rate Falacy

Conditional probabilities

We write

$$Pr(HasDisease = y \cap FailsTest = y)$$

to mean "the probability that $HasDisease = y$ *and* $FailsTest = y$ ".
For *conditional probabilities* we write

$$Pr(HasDisease = y | FailsTest = y)$$

to mean "the probability that $HasDisease =$y *given* $FailsTest = y$ ".
This can be written precisely as

$$Pr(HasDisease = y | FailsTest = y) = \frac{Pr(HasDisease = y \cap FailsTest = yes)}{Pr(FailsTest = y)}$$

WPI

Conditional probabilities

Note, these are very different things. For example, when $Pr(HasDisease = y)$ is independent of $Pr(FailsTest = y)$ we have that

$$Pr(HasDisease = y \cap FailsTest = y) = Pr(HasDisease = y)Pr(FailsTest = y)$$

but, still assuming independence,

$$\begin{aligned} Pr(HasDisease = y | FailsTest = y) &= \frac{Pr(HasDisease = y \cap FailsTest = y)}{Pr(FailsTest = y)} \\ &= Pr(HasDisease = y) \end{aligned}$$

Bayes Theorem, Law of Total Probability, and Base Rate Fallacy

Bayes theorem

$$Pr(HasDisease = y|FailsTest = y) = \frac{Pr(FailsTest = y|HasDisease = y)Pr(HasDisease = y)}{Pr(FailsTest = y)}$$

Law of total probability

$$\begin{aligned} Pr(FailsTest = y) &= Pr(FailsTest = y|HasDisease = y)Pr(HasDisease = y)+ \\ &\quad Pr(FailsTest = y|HasDisease = n)Pr(HasDisease = n) \\ &= Pr(FailsTest = y|HasDisease = y)Pr(HasDisease = y)+ \\ &\quad Pr(FailsTest = y|HasDisease = n)(1 - Pr(HasDisease = y)) \end{aligned}$$

WPI

Things to think about

- If you are given a table with the appropriate probabilities, can you compute the equations below?
- In particular, can you do the calculation in the iPython notebook by hand?

Bayes theorem

$$Pr(HasDisease = y|FailsTest = y) = \frac{Pr(FailsTest = y|HasDisease = y)Pr(HasDisease = y)}{Pr(FailsTest = y)}$$

Law of total probability

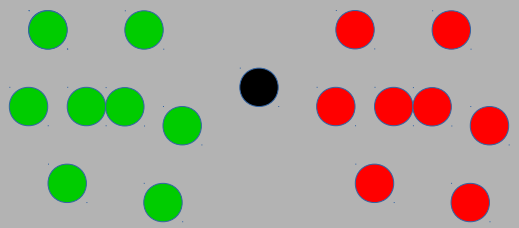
$$\begin{aligned} Pr(FailsTest = y) &= Pr(FailsTest = y|HasDisease = y)Pr(HasDisease = y)+ \\ &\quad Pr(FailsTest = y|HasDisease = n)Pr(HasDisease = n) \\ &= Pr(FailsTest = y|HasDisease = y)Pr(HasDisease = y)+ \\ &\quad Pr(FailsTest = y|HasDisease = n)(1 - Pr(HasDisease = y)) \end{aligned}$$

WPI

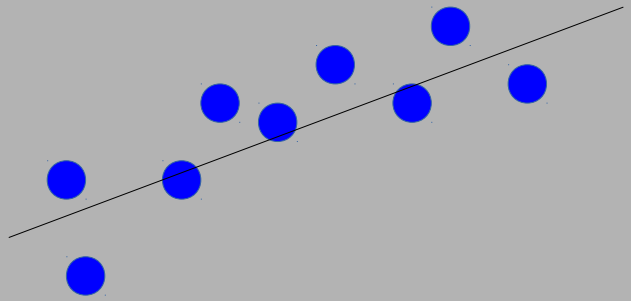
Class 6
Machine Learning

We have four days we will cover four topics.

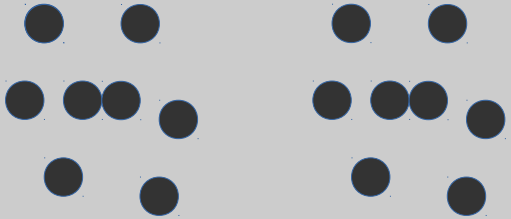
Supervised Classification



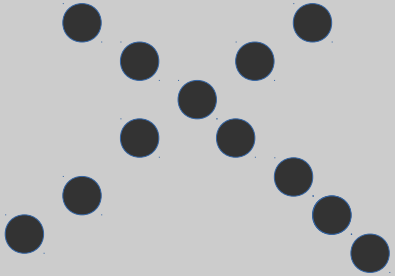
Supervised Regression



Unsupervised Clustering



Dimension Reduction



WPI

What is PCA?

- Principle Component Analysis
  - Commonly used tool for visualization and data pre-processing.

What is Linear Support Vector Machine (SVM)?

- Maximum margin classifier
  - Computes a linear “decision boundary” that splits the data into two regions.
  - Allows one to predict a classification of a point based upon which side of the decision boundary it lay on.

What is K-NN?

- K-nearest neighbors
- Another common classification algorithm
  - Perhaps the most common