Big Idea #3

Data and information facilitate the creation of knowledge.

- People use computer programs to process information to gain insight and knowledge.
- Computing facilitates exploration and the discovery of connections in information.
- Computational manipulation of information requires consideration of representation, storage, security, and transmission.
Stanford researchers produce first complete computer model of an organism

A mammoth effort has produced a complete computational model of the bacterium Mycoplasma genitalium, opening the door for biological computer-aided design.
• Example Problems
• Challenges
• Big Data *Unplugged*
• Paradigms
• *Hands-on* Visualization & Data Mining
Example: Internet Search

- Enormous amounts of content on the Internet

47 billion  17 billion  3.3 billion

- Seek relevant results in less than a second
Example: Internet Search

Prior to searches (happens continuously):
1. Crawl the web to locate pages
2. Create index of pages

For each search (in fraction of a second):
1. Locate pages with keywords
2. Rank pages by relevance
3. Return results to user
Example: Climate Analysis

• Analyze current and historical weather data
  – Sensor readings from 1000s of locations
  – Satellite/radar images
  – Geographic features

• Visualize predictions for many audiences
Example: Netflix Recommendations

- Recommend movies from Netflix’s collection

- Accuracy of predictions impacts subscriptions
Example: Netflix Recommendations

• Many factors can influence viewing behavior
  – Movie characteristics: cast, year, genre, duration
  – Personal history: movies watched, queue
  – Social: ratings, reviews

• Recommendations include categories and movies, presented in a specific order
Challenge: Collection

Where does the data come from?

• Input from humans, instruments/sensors, existing datasets, etc.
• Potentially many sources
• Transport data from source to repository
Challenge: Organization

How is the data structured?

• Data needs to be labeled, sorted, etc.
• Relationships may exist between pieces
• Exclude inaccurate or unknown data
Challenge: Storage

*How do we store large volumes of data?*

- Need space for 100s of Terabytes of data (modern hard drive holds 1 TB)
- Data needs to be *efficiently* accessed by servers doing computation
Challenge: Computation

How is the data processed to obtain desired information?

• Algorithms determine actions to perform
• Need computers to run the algorithms
• May be constrained by time, space, etc.
Challenge: Visualization

How is the data (or results) presented?

• Seek clear, concise representation of the data
• Emphasize desired information
• May require many related visualizations
Big Data Unplugged

• Word count
  – Conceptually simple
  – Relevant for Internet search

• Count how many times each unique word occurs

• Want *speed* and *accuracy*
Big Data Unplugged

- Who held what data?
- How was data passed?
- What algorithm did each person execute?
- How was the final result obtained?
- How did you present the final result?
Paradigm: MapReduce

- Leverage parallelization
- Divide analysis into two parts
  - Map task: given a subset of the data; extract relevant data and obtain partial results
  - Reduce task: receive partial results from each map task; combine into final result
Paradigm: MapReduce

• Used for Internet search
  – Map task: given a part of the index; identify pages containing keywords and calculate relevance
  – Reduce task: rank pages based on relevance

• Infrastructure requirements
  – Many machines to run map tasks in parallel
  – Ability to retrieve and store data
  – Coordination of who does what
Paradigm: Cloud Computing

- Large collections of processing and storage resources used on demand
- Sell resources (machines, GB of storage, etc.) for some period of time
Paradigm: Cloud Computing

- Infrastructure-as-a-service
- Platform-as-a-service
- Storage-as-a-service
Paradigm: Cloud Computing

• Benefits for *users*
  – Only pay for what you use
    100 servers at $1/hour for 1 hour = $100
    1 server at $1/hour for 100 hours = $100
  – Externally managed

• Benefits for *cloud providers*
  – Economies of scale (space, equipment, etc.)
Paradigm: Data Mining

- Identify patterns and relationships in data
- Used to rank, categorize, etc.
- Commonly associated with artificial intelligence and machine learning
Paradigm: Data Mining

• Categorization algorithms
  – Rules > ZeroR: pick most common
  – Trees > J48: decision tree
  – Bayes > NaiveBayes: based on probabilities

• Clustering algorithms
Paradigm: Visualization

• Wide array of ways to view data (or results)
  – Conventional: line, bar, pie charts
  – Alternative: bubble chart, tree map, world map
  – Text: tag cloud, word tree
Hands-On

• Data Mining in Weka
  – Computer > cshs2012 (Z:) > launch_weka
  – Data in Z:/datasets
  – Rules > ZeroR, Trees > J48, Bayes > NaiveBayes

• Visualization using Many Eyes
  – Search for “one fish” datasets or play with any dataset
Resources

- ManyEyes (http://www-958.ibm.com)
- Weka (http://www.cs.waikato.ac.nz/ml/weka)
- Datasets (http://archive.ics.uci.edu/ml/)
- Google Insights for Search (http://www.google.com/insights/search)
- WebMapReduce (http://webmapreduce.sourceforge.net/)
- Amazon Web Services in Education (http://aws.amazon.com/education/)