HW1 PROBLEM 1: Big Data Characteristics (50%)

(a) Access Log Data:
   1) Amount: huge. As what is shown in the example, almost every minute has one or two new logs. So the amount of logs over 5 years is huge.
   2) Size: small. The size of each data point is less than 1 line.
   3) Infinity: yes. The logs are keep arriving. The velocity is 1-2 logs per minute, which means data arrives faster than it can be processed.
   4) Structure: semi structured. The logs are not tables but they have fixed patterns. They all consist of time, IP address and other details.
   5) Complexity: no. The logs are explicit.

(b) Wikipedia articles:
   1) Amount: huge. The number of Wikipedia articles is more than 43millions.
   2) Size: small. Most sizes of text-only articles are smaller than 1MB.
   3) Infinity: yes. The Wikipedia articles change faster than it can be processed..
   4) Structure: unstructured. The articles do not have a fixed format.
   5) Complexity: yes. There are relationships between some articles.

(c) Chemical compounds:
   1) Amount: huge. The number of chemical compounds in the world is huge.
   2) Size: small. The example data set only has text so its size is small.
   3) Infinity: no. There is not a rapid change of chemical compounds.
   4) Structure: structured. The data sets are represented as tables and they have fixed patterns.
   5) Complexity: yes. Many chemical compounds are inter-related.

(d) Both data sets described in (a) and (b) are represented as text.
   1) The main difference is that data sets in (a) have fixed formats while data sets in (b) doesn't. Many data sets in (b) depend on others, which means they have complexity.
   2) We have to know what does each data set mean and build some search index or tables for them.

(e) For the example data sets described in (a)-(c), what are the data points and what data types/data structures do we use to represent those data points programmatically?
   For (a), the data point is represented as text.
   For (b), the data point is represented as text.
   For (c), the data point is represented as table or array.

HW1 PROBLEM 2: Bonferroni's Principle (25%)

The probability of any p people all deciding to visit a hotel on any given day is 10^{-2p}.

The probability that they will visit the same hotel is this probability times $(10^{-5})^{p-1}$,

Thus, the chance that they will visit the same hotel on one given day is $10^{-2p-5(p-1)}$

The chance that they will visit the same hotel on two different given days is powers of d of this number. It is $10^{[-2p-5(p-1)]d}$.

The number of pairs of people is $\binom{10^9}{p} = \frac{10^{9p}}{p!}$. The number of pairs of days is $\binom{10^3}{d} = \frac{10^{3d}}{d!}$.

The expected number of events that look like evil-doing is the product of the number of pairs of people, the number of pairs of days, and the probability that any one pair of people and pair of days is an instance of the behavior we are looking for. That number is

$$\binom{10^9}{p} \times \binom{10^3}{d} \times [(10^{-2})^P \times 10^{-5(p-1)}]^d = \frac{10^{9p}}{p!} \times \frac{10^{3d}}{d!} \times 10^{[-2p-5(p-1)]d} = \frac{10^{9p+8d-7pd}}{p!d!}$$

HW1 PROBLEM 3: The Unreasonable Effectiveness of Data (25%)

(a) What are the differences between unlabeled and labeled/annotated data?Unlabeled data it's unfiltered and contains incomplete data and errors. It has large scale. While labeled data is formatted. It is annotated with carefully hand corrected part of speech tags. But sometimes it is not available.

(b) Summarize the data-based approach described in the article
Semantic Interpretation: First, we can extract a set of schemata from the tables' column labels of given corpus. Then examine the co-occurrences of attribute names in these schemata. A set of attributes have the same meaning if they rarely show up together but always occur with the same other attributes. This hypothesis can be verified if data elements have a significant overlap. What is more, a schema autocomplete feature can be offered for database designers. Then we can automatically combine data from multiple tables in this collection. We can also combine data from multiple tables with data from other sources.

(c) What are the limits of this approach?
There is still a scientific problem of interpreting the content, which is mainly that of learning as much as possible about the context of the content to correctly disambiguate it.
The same meaning can be expressed in many different ways, and the same expression can express many different meanings. We can't expect to have an ontology for every possible value of this attribute.
Large scale data sets are required.