Basics and Big data

Big Data: 4 Vs

--What are they and what are some applications?

Analytics: some broad views

- From Oxford Dictionary:
  - *the **systematic** computational analysis of **data or statistics***
  - *information resulting from the **systematic** analysis of **data or statistics***
- From Techopedia:
  - *Analytics is the **scientific process** of discovering and **communicating** the meaningful patterns which can be found in data. It is concerned with **turning raw data into insight for making better decisions**. Analytics relies on the application of statistics, computer programming, and operations research in order to quantify and gain insight to the meanings of data. It is especially useful in areas which record a lot of data or information.*
- From 'Innovating with Analytics', MIT Sloan Management Review:
  - *"...there is a strong correlation between driving competitive advantage and innovation with analytics and a company's effectiveness at managing the **information transformation cycle**, that is: **capturing data, analyzing information, aggregating and integrating data, using insights to guide future strategy and disseminating information and insights**."*

MapReduce

'hello world': word counting in large corpus

```
map(String key, String value):
  // key: document name
  // value: document contents
  for each word w in value:
    EmitIntermediate(w, "1");

reduce(String key, Iterator values):
  // key: a word
  // values: a list of counts
  int result = 0;
  for each v in values:
    result += ParseInt(v);
  Emit(AsString(result));
```

K-Means in MapReduce (more complex example)

k-means::Map
Input: Data points D, number of clusters k and centroids
1: for each data point d Є D do
2: Assign d to the closest centroid
Output: centroids with associated data points

---

k-means::Reduce
Input: Centroids with associated data points
1: Compute the new centroids by calculating the average of data points in cluster
2: Write the global centroids to the disk
Output: New centroids

Learning and Inference

What is a good definition of learning (hint: think of Herb Simon's definition)?

What is a critical feature of learning?

What are some similarities and differences between inference, reasoning and prediction?

Naïve Bayes and Bayesian Networks

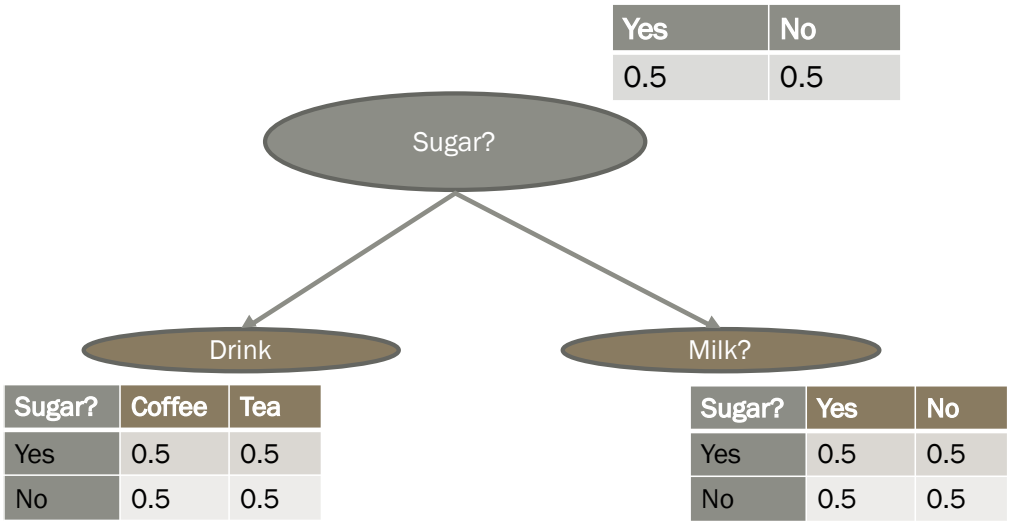
x1	x2	Y
A1: drink	A2: milk?	C: sugar?
coffee	no	yes
coffee	yes	no
tea	yes	yes
tea	no	no

Can you train a Naïve Bayes classifier to predict whether the customer wants sugar or not?

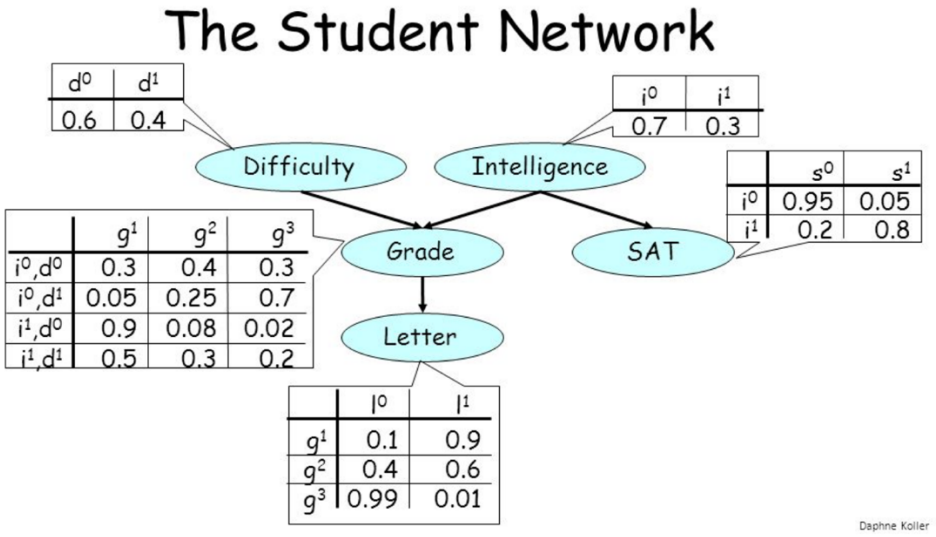
What is P(coffee | sugar)?

How would you get the Naïve Bayes parameters?

Use maximum likelihood:



More complex example (Bayesian networks)



Given the above, what is the probability that I get a good letter (l1) given grade g3 and SAT score s1? Show your results.

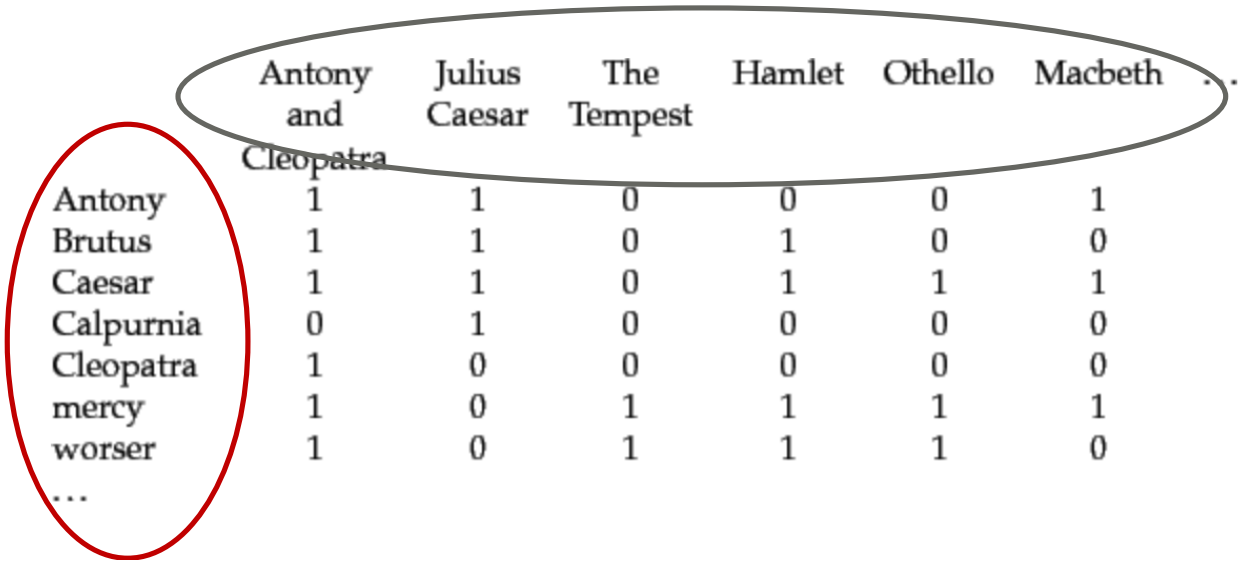
Is Bayesian network a generative or discriminative model? What are the main differences between the two types of models?

Supervised vs. unsupervised text classification

What are some examples and applications of supervised vs. unsupervised text classification?

Given training data, are there cases you can think of where you would still want to use unsupervised, rather than supervised, text classification?

Information retrieval basics: incidence matrix and tf-idf



	Antony and Cleopatra	Julius Caesar	The Tempest	Hamlet	Othello	Macbeth	. .
Antony	1	1	0	0	0	1	
Brutus	1	1	0	1	0	0	
Caesar	1	1	0	1	1	1	
Calpurnia	0	1	0	0	0	0	
Cleopatra	1	0	0	0	0	0	
mercy	1	0	1	1	1	1	
worser	1	0	1	1	1	0	
. . .							

Given a vocabulary (number of unique words in corpus) of size V and D documents, what is the number of elements in the matrix?

Given a set of words ('query') how could you use a matrix such as this for rudimentary information retrieval?

What does tf-idf stand for? Is it fair to say that the 0-1 incidence matrix is a simplified version of tf-idf?

What are some advantages of tf-idf that a 0-1 incidence matrix does not have?

Give a hypothetical concrete example where IR works just as well with a 0-1 incidence matrix as tf-idf. Next, give an example where it makes a clear difference.

Consider the question about the V words and D documents above. How many **bytes** would you need for an incidence matrix? What about tf-idf? Can you think of applications or devices where this difference could matter tremendously?