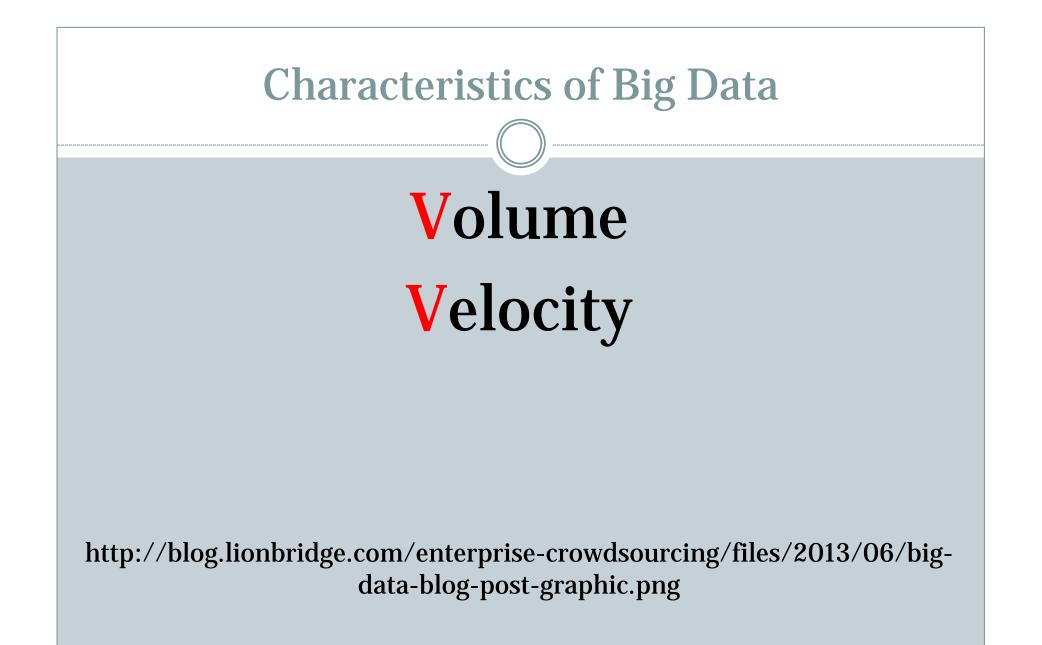


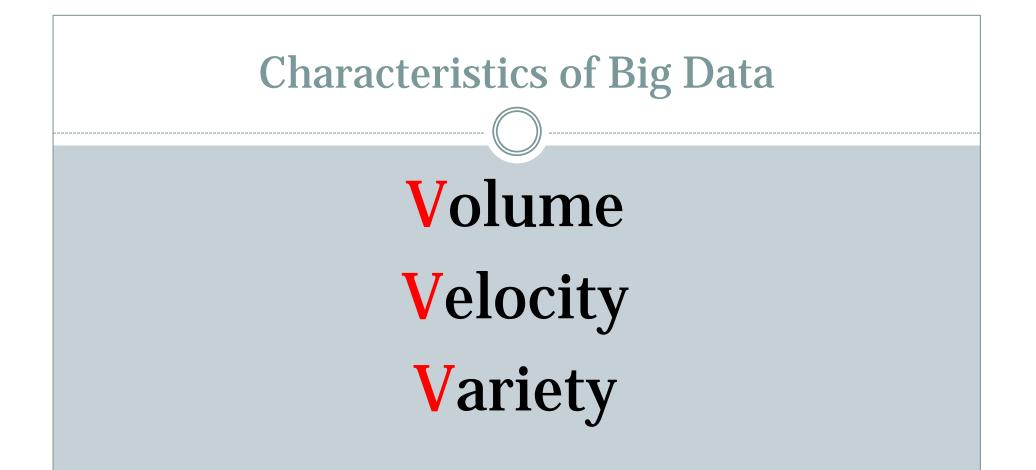
# What is Big Data?

- Increasingly popular search query
  - o http://www.google.com/trends/explore#q=big%20data

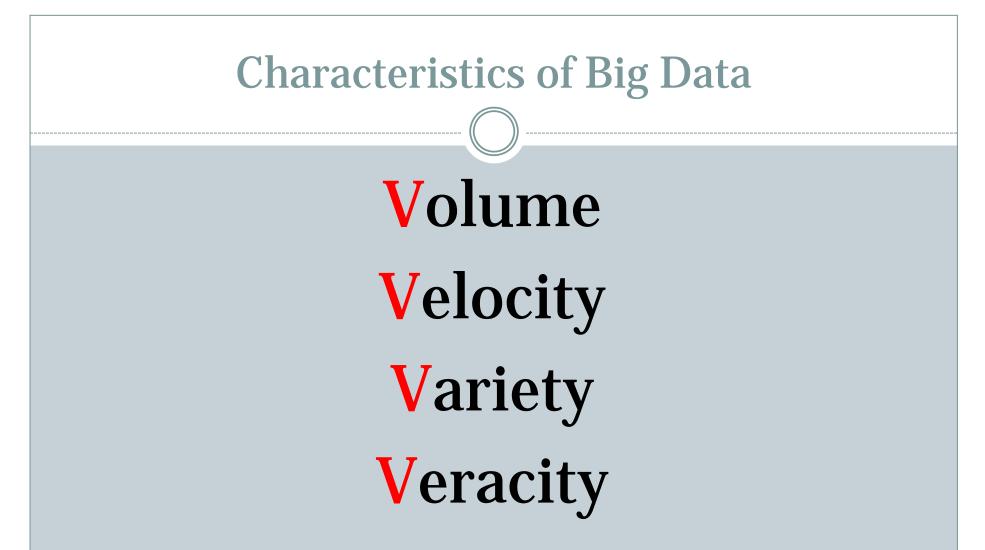








• http://www.datasciencecentral.com/forum/topics/the-3vs-that-define-big-data



- http://www.sec.gov/Archives/edgar/data/51143/000110465913015636/g61551bii004.gif
- http://www.datasciencecentral.com/profiles/blogs/data-veracity
- http://www.ibmbigdatahub.com/sites/default/files/infographic\_file/4-Vs-of-big-data.jpg



#### What to do with Big Data?

#### Advances in Computer Science (CS)

• sample companies

#### • sample applications in **analytics**

× Focus of speakers

#### • EMC

o http://www.emc.com/campaign/bigdata/index.htm

• EMC (data storage hardware)

o <u>http://www.emc.com/campaign/bigdata/index.htm</u>

- EMC (data storage hardware)
  - o <u>http://www.emc.com/campaign/bigdata/index.htm</u>
- Oracle
  - o http://www.oracle.com/us/technologies/big-data/index.html

- EMC (data storage hardware)
  - o <u>http://www.emc.com/campaign/bigdata/index.htm</u>
- Oracle (database software)
  - o http://www.oracle.com/us/technologies/big-data/index.html

- EMC (data storage hardware)
  - o <u>http://www.emc.com/campaign/bigdata/index.htm</u>
- Oracle (database software)
  - o http://www.oracle.com/us/technologies/big-data/index.html
- SAS
  - o <u>http://www.sas.com/en\_us/insights/big-data.html</u>

- EMC (data storage hardware)
  - o <u>http://www.emc.com/campaign/bigdata/index.htm</u>
- Oracle (database software)
  - o http://www.oracle.com/us/technologies/big-data/index.html
- SAS (analytics software)
  - o http://www.sas.com/en\_us/insights/big-data.html

- EMC (data storage hardware)
  - o <u>http://www.emc.com/campaign/bigdata/index.htm</u>
- Oracle (database software)
  - o http://www.oracle.com/us/technologies/big-data/index.html
- SAS (analytics software)
  - o <u>http://www.sas.com/en\_us/insights/big-data.html</u>
- IBM
  - o http://www.ibm.com/big-data/

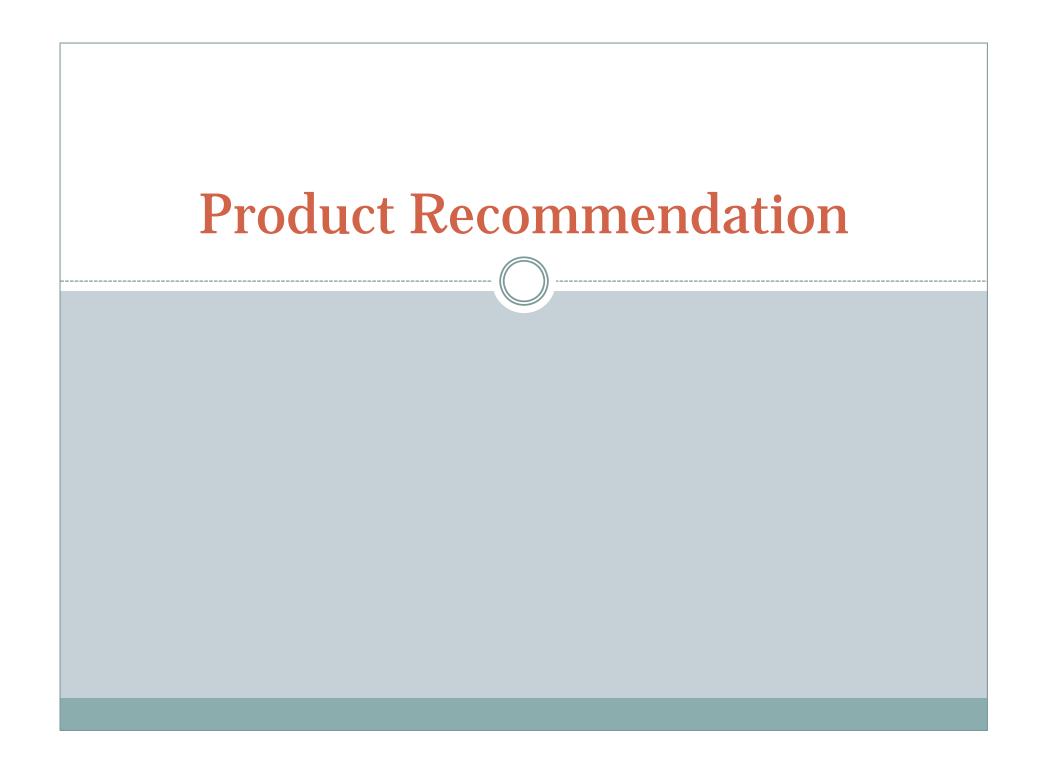
- EMC (data storage hardware)
  - o <u>http://www.emc.com/campaign/bigdata/index.htm</u>
- Oracle (database software)
  - o http://www.oracle.com/us/technologies/big-data/index.html
- SAS (analytics software)
  - o <u>http://www.sas.com/en\_us/insights/big-data.html</u>
- IBM (computational hardware, database & analytics software, services)
  - o http://www.ibm.com/big-data/

### **Sample Applications in Analytics**

#### Automated

• Recommendation of products (amazon, netflix)

• Organization of news articles (google news)



### **Recommendation Systems**

#### • Netflix

• Recommending movies

#### Amazon

• Recommending products of many kinds

### **Netflix Prize**

• netflixprize.com

#### • \$1 million

- 10% improvement in accuracy over Netflix algorithm/system (in 5 years)
- o 51K contestants, 41K teams, 186 countries
- Did any team win the prize?
  - × If so, how long did it take?

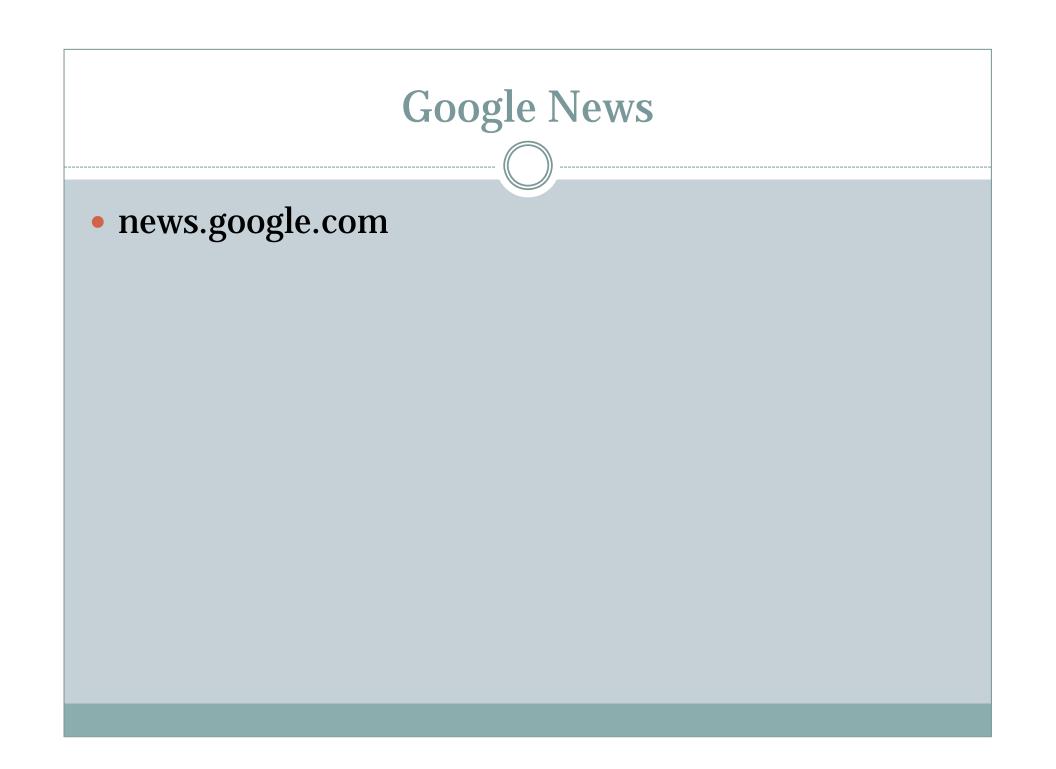
#### Netflix Prize Data (1998-2005)

- Customers
  - o 480,189 (ID: 1 − 2,649,429)
- Movies
  - o 17,770 (ID: 1−17,770)
  - o ID, title, year
- Ratings given in Training Set
  - 100,480,507
  - o min=1; max=17,653; avg=209 ratings per customer
  - o Rating scale: 1 − 5
  - o Date
- Ratings to predict in Qualifying Set
  2,817,131
- About 1 GB (700 MB compressed)

Netflix Problem									
	M1	M2	M3	M4	M5	M6	M7	M8	M9
C1	?		1		3		4	?	
C2	2		3		1		4	5	
C3	2	5	3	3	3		4	4	1
C4	2		3					2	
C5						4	1	3	3

? = unknown rating to be predicted





# **Google News**

- news.google.com
  - News organizer from many sources
  - Krishna Bharat, 2006 http://googleblog.blogspot.com/2006/01/and-now-news.html
    - "using computers to organize the world's news in real time and providing a bird's eye view of what's being reported on virtually any topic.

# **Google News**

- news.google.com
  - News organizer from many sources
  - Krishna Bharat, 2006 http://googleblog.blogspot.com/2006/01/and-now-news.html
    - "using computers to organize the world's news in real time and providing a bird's eye view of what's being reported on virtually any topic.
    - Sy presenting news "clusters" (related articles in a group), we thought it would encourage readers to get a broader perspective by digging deeper into the news -- reading ten articles instead of one, perhaps -- and then gain a better understanding of the issues, which could ultimately benefit society"

## **Google News Problem**

- Input
  - News articles
- Output
  - Clusters of news articles
- Algorithm (how)
  - **o** ?

### Animation of a Clustering Algorithm

http://shabal.in/visuals/kmeans/1.html

#### **Sample Applications in Analytics**

#### Automated

• Spam Detection -- Ryan Stansifer

**o** DNA Sequence Analysis -- Debasis Mitra

#### **Upcoming CS Outreach Events**

Mar: teacher workshop

#### Mar-Apr: student competition

- Boulder Dash
- Details in mid March
- Jul: summer camps

cs.fit.edu/~pkc/cs4hs

**Questions?**