Big Data: A Closer Look

What is Big Data?

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

What is Big Data?

the V's

Characteristics of Big Data

Volume

http://www.businessweek.com/articles/2014-03-06/interactive-graphic-how-big-is-big-data

Characteristics of Big Data

Volume

Velocity

http://blog.lionbridge.com/enterprise-crowdsourcing/files/2013/06/big-data-blog-post-graphic.png

Characteristics of Big Data

Volume

Velocity

Variety

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

Characteristics of Big Data

Volume

Velocity

Variety

Veracity

- 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?

What to do with Big Data?

- Advances in Computer Science (CS)
  - sample companies
  - sample applications in analytics
    - Focus of speakers

Sample Companies

- EMC
  - http://www.emc.com/campaign/bigdata/index.htm

Sample Companies

- EMC (data storage hardware)
  - http://www.emc.com/campaign/bigdata/index.htm

Sample Companies

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

Sample Companies

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

Sample Companies

- EMC (data storage hardware)
  - http://www.emc.com/campaign/bigdata/index.htm
- Oracle (database software)
  - http://www.oracle.com/us/technologies/big-data/index.html
- SAS
  - http://www.sas.com/en_us/insights/big-data.html

Sample Companies

- EMC (data storage hardware)
  - http://www.emc.com/campaign/bigdata/index.htm
- Oracle (database software)
  - http://www.oracle.com/us/technologies/big-data/index.html
- SAS (analytics software)
  - http://www.sas.com/en_us/insights/big-data.html

Sample Companies

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

Sample Companies

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

Sample Applications in Analytics

- Automated
  - Recommendation of products (amazon, netflix)
  - Organization of news articles (google news)

Product Recommendation

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)
  - 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
  - 480,189 (ID: 1 – 2,649,429)
- Movies
  - 17,770 (ID: 1 – 17,770)
  - ID, title, year
- Ratings given in Training Set
  - 100,480,507
  - min=1; max=17,653; avg=209 ratings per customer
  - Rating scale: 1 – 5
  - 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

Organizing News Articles

Google News

- news.google.com

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.
    - By 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)
  - ?

Animation of a Clustering Algorithm

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

Sample Applications in Analytics

- Automated
  - Spam Detection -- Ryan Stansifer
  - 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?