Term Report CSS499, Summer 2016
Pichsendarong Leng

Application	**Computational Intensiveness**	**Applicable with Agent-based System**
Image Recognition • edge detection	High	Applicable
Are liquor Stores too close to schools, libraries, and parks? -Problem: increase in complaints form citizens about nuisance activity ear liquor stores, including public drunkenness, disturbing the peace	Medium	Applicable
United States Tornadoes It reveals how the use of spatial analysis helps reveal patterns in data	Low	Applicable
Analyzing Crime – Detect Patterns City councils need to be determine whether or not liquor stores influence crime activity	Low	Applicable
Historic Croplands Dataset The size of the data that s available is small	Low	Applicable
Global Agriculture Land The Global Pastures data set represents the proportion of land areas used pasture land in year 2000. The data set is too small-not an ideal application	Low	Applicable
Gridded Application of the World This is a gridded data product that renders global population data at the scale and extent required to demonstrate the spatial relationship of human populations and the environment across the globe	Low	Applicable

Categorization of Big Data Applications

Category	Application	Data structure	Algorithm	Computational Intensive	Spatial Requirement	Applicable with Agents
S, G, Mr, Mclass	Synopsys	Array, network or anything else?	A brief explanation. Parallelizable?	Complete in seconds, minutes, or hours?	MB, GB, or TB?	If so how? If not, why?
S	Indexing of Astronomica	Network	-Read previous data	Minutes	GB or TB	Applicable because

	l Objects		-analyze the image -perform the calculation			each objects has its own coordinate
G	Predicting Generator Failure	Networ k	-Read data -predict which generator will likely to fail -how failure affect the generator netowrk	Seconds	GB	Applicable —each generator can have its own place and are connected to other generators

Background of "Indexing of Astronomical Objects" (application mentioned in table above):
When astronomers analyze telescope images, they match the observed objects to the catalog that contains with billions of objects.

Identification of Big Data Applications
My approach is to get started with exploring what applications MapReduce, Spark, and Storm have already parallelized.

Categorization of Big Data Applications

Category	Application	Data structure	Algorithm	Computational Intensive	Spatial Requirement	Applicable with Agents
S, G, Mr, Mclass	Synopsys	Array, network or anything else?	A brief explanation. Parallelizable?	Complete in seconds, minutes, or hours?	MB, GB, or TB?	If so how? If not, why?

Uber Trip Analysis
Category: GIS
Possible Application:

- Explore which neighborhood has the most rides (the one in the article)
- Find the busiest time of the day and week
- Average duration of a trip

Data Structure: Array is suitable because most applications only involve counting.
Algorithm: Coordinates given by Uber needs to be converted to a coordinate system that matches the neighborhood dataset. Find the starting point and destination point of each trip, and group trip that has the same neighborhood accordingly.
Computational Intensive: Minutes
Spatial Requirement: MB or GB are good sizes, but the bigger the size the more in depth analysis we can get out of.

Applicable with Agents: Applicable with our agents. In MASS, we need to distribute the data to each node and have them find the starting and destination point of each trip, and group them accordingly. Before ending the computation, each node needs to merge their count and perform a sum.

GeoTrellis Transit
Category: GIS
Possible Application: GeoTrellis is actually geoprocessing engine that already provides REST API to answer questions how far can you travel from a certain location with a given amount of time and etc.
Building application that is similar to this with our system, MASS, would require a lot of data, time, and knowledge of GIS.

Flight Application
Category: GIS
Possible Application:

- Estimate flight delays
- Explore which airport serves the most flights
- Explore which airline serves the most flights
- Another possible application is to find which type of weather causes flight delay by combining flight datasets and weather datasets (which I haven't found available data yet)

Data Structure: Array is suitable.
Algorithm: I don't have an exact algorithm for this one because I haven't chosen an application yet. What we do have right now is dataset whose size is more than 12gb.
Computational Intensive: Possibly hours
Spatial Requirement: The size of the data that is available is 12gb.
Applicable with Agents: Applicable with our agents because we just use them to process information.

airport, airline, and routes data: http://openflights.org/data.html
flight data: http://stat-computing.org/dataexpo/2009/the-data.html

Categorization of Big Data Applications

Category	Application	Data structure	Algorithm	Computational Intensive	Spatial Requirement	Applicable with Agents
S, G, Mr, Mclass	Synopsys	Array, network or anything else?	A brief explanation. Parallelizable?	Complete in seconds, minutes, or hours?	MB, GB, or TB?	If so how? If not, why?

R-Tree: is one of the data structures used for storing spatial objects or multidimensional data (polygons, geo coordinates). It has a very similar data behavior to B-Tree. This data structure is mostly used in database management system for spatial query processing.

I haven't found specific data analysis that uses R-Tree though I have seen authors mention that "R-Tree is used in many applications." My theory is that as long as an analysis relates to multidimensional data (like geographical coordinates), it can be used with R-Tree data structure. Below, there are two of the links that I read about R-Tree data structure.

http://www.cse.cuhk.edu.hk/~taoyf/course/wst501/notes/lec13.pdf
http://delab.csd.auth.gr/~apostol/pubs/rtrees03.pdf

Netflix Film Rating
Netflix provided a dataset containing 100 million ratings (includes 17,000 movies and 480,000 customers) from 1998 to 2005. The purpose was to find out improve the accuracy of its previous rating prediction algorithm.
Category: M or Neural Network
Application: Predict user ratings for films based on previous data including movie id, user id, rating, title, year, and date.
Data Structure: Graph is suitable
Algorithm: I'm not confident on how to start with the algorithm.
Computational Intensive: Since the training data size provided is about 750mb, computation time might be minutes.
Spatial Requirement: There is not spatial requirements.
Applicable with Agents: I have little background in Machine Learning and Artificial neural networks but I have to say it might be possible based on reading articles and watching videos. The main function of the application is to predict user rating. Artificial neural network deals with graphs structure. The purpose is to improve the accuracy of the prediction through figuring out what values suit the best for edges in the graph by using the training data set. In MASS, we can have each node process small chunk part of the training dataset and communicate with main node to update and ask for values of edges in the graph. *(What I am saying here is just a thought based on my reading articles, not in depth).*

https://en.wikipedia.org/wiki/Netflix_Prize
http://academictorrents.com/details/9b13183dc4d60676b773c9e2cd6de5e5542cee9a (dataset)

Flight Application *(from previous report)*
Category: GIS
Possible Application:

- Explore which part/airport of the world serves the most flight during a certain period of time

Data Structure: Array is suitable since most of the computation will just be processing information.
Algorithm: Distribute data to all available nodes and have them go through the data and count the flight information. Two possible approaches:

- Have each node be responsible for all of airports around the world (which is a lot) and increase the count of an airport accordingly as they go through the dataset. Finally, have them sum up the count at the end of the computation. OR
- Each node is responsible for a certain number of airports. Whenever a node encounters an airport that it doesn't know it needs to let the node that is responsible for that airport to increase the count.

In my opinion, the first approach is way better with less communication.
Computational Intensive: Possibly hours since the data is really big
Spatial Requirement: The size of the data that is available is 12gb.
Applicable with Agents: Applicable with our agents because we just use them to process information.

Category
S: Scientific Computing
G: GIS or Spatial Data Analysis
M: Machine Learning
Mr: Regression
Mca: Classification
Mcu: Clustering
Mar: Association rule
Man: Artificial neural networks