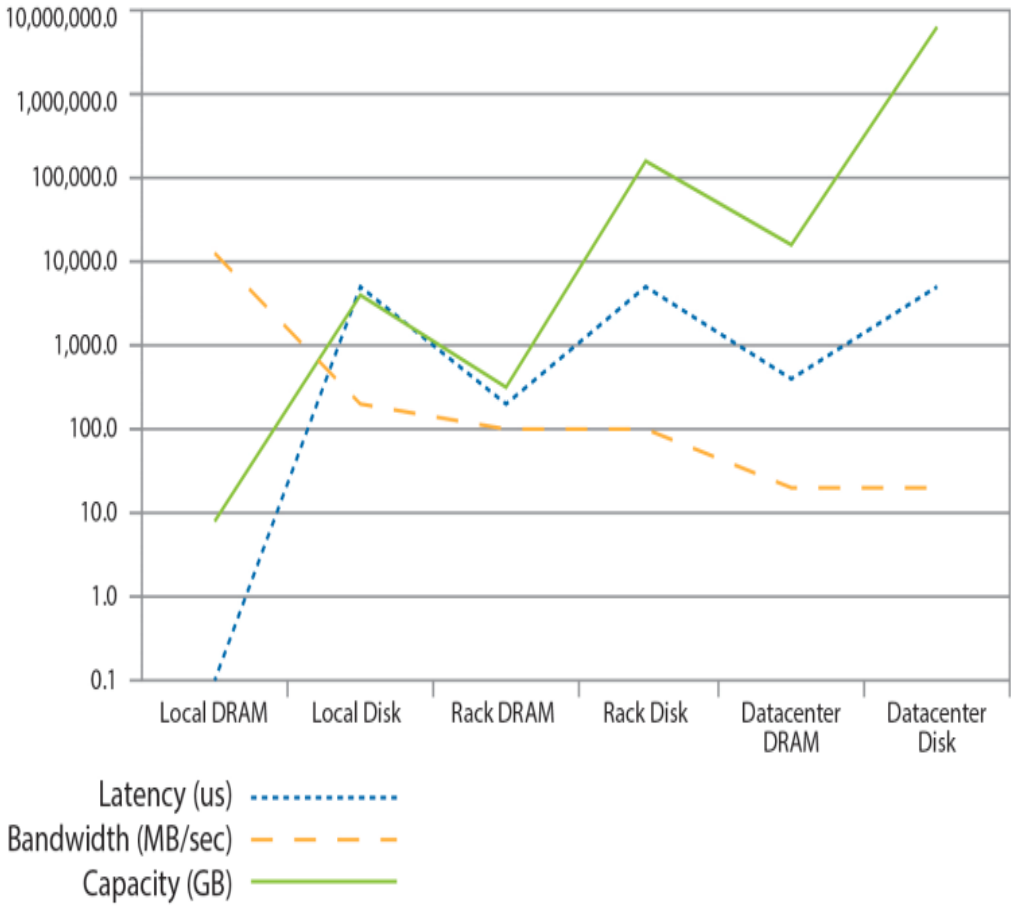
Tutorial 1: Introduction to Data Science

Li Yuan

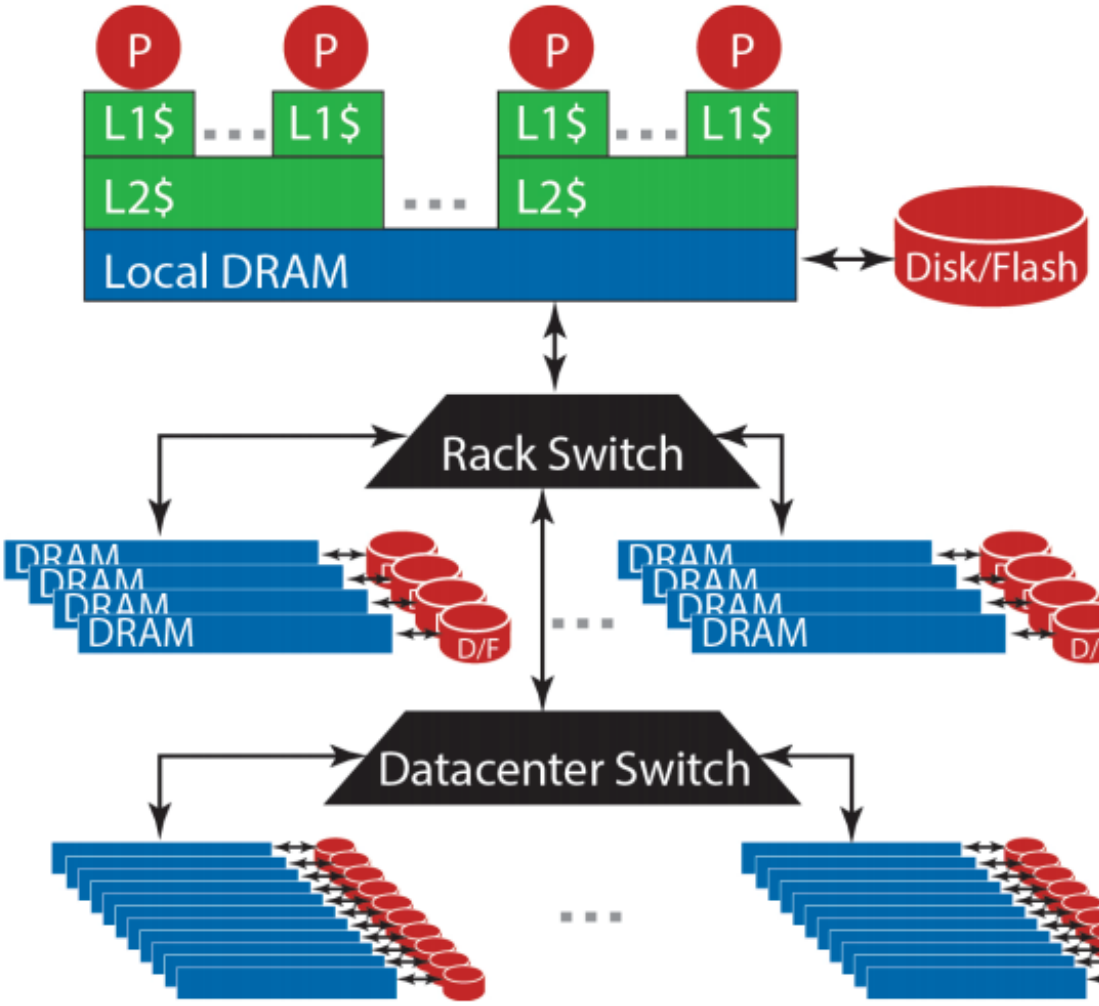
li.yuan@u.nus.edu

Question 1a

Why the latency of Rack DRAM is much higher than that of Local DRAM?



Storage Hierarchy



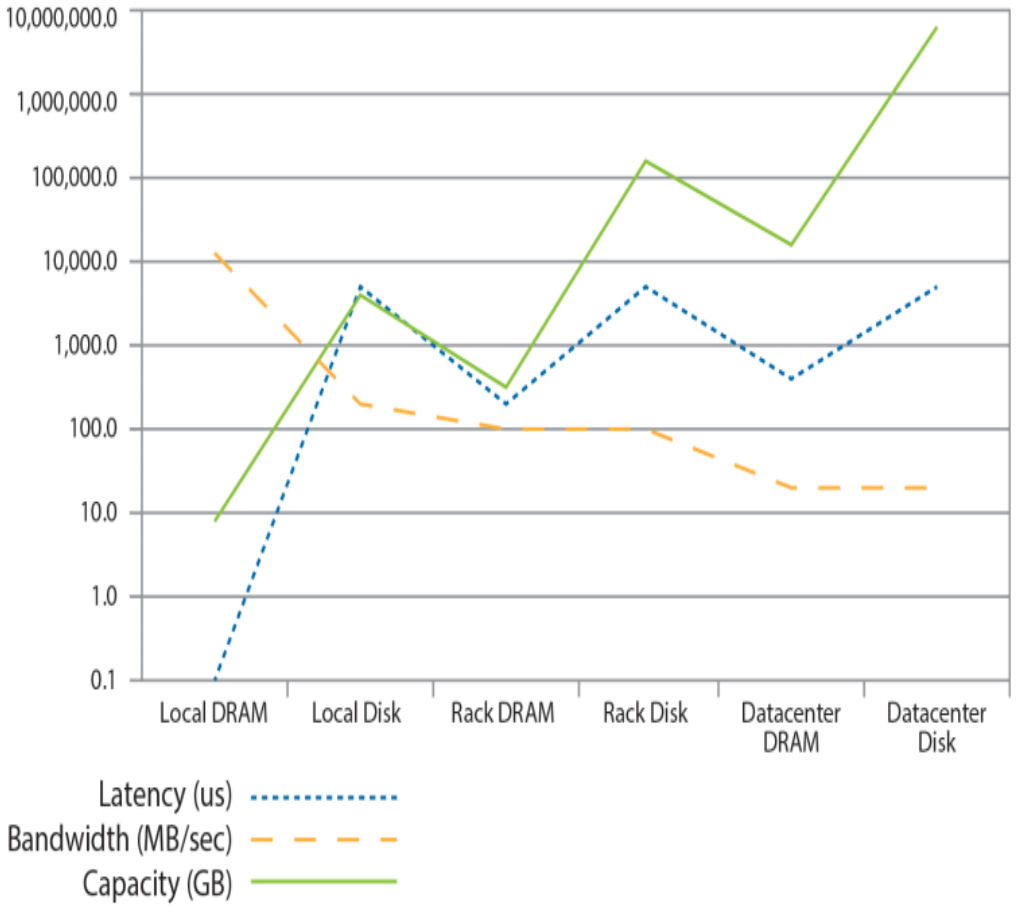
One Server
DRAM: 16 GB, 100 ns, 20 GB/s
Disk: 2T B, 10 ms, 200 MB/s
Flash: 128 GB, 100 us, 1 GB/s

Local Rack (80 servers)
DRAM: 1 TB, 300 us, 100 MB/s
Disk: 160 TB, 11 ms, 100 MB/s
Flash: 20 TB, 400 us, 100 MB/s

Cluster (30 racks)
DRAM: 30 TB, 500 us, 10 MB/s
Disk: 4.80 PB, 12 ms, 10 MB/s
Flash: 600 TB, 600 us, 10 MB/s

Question 1a

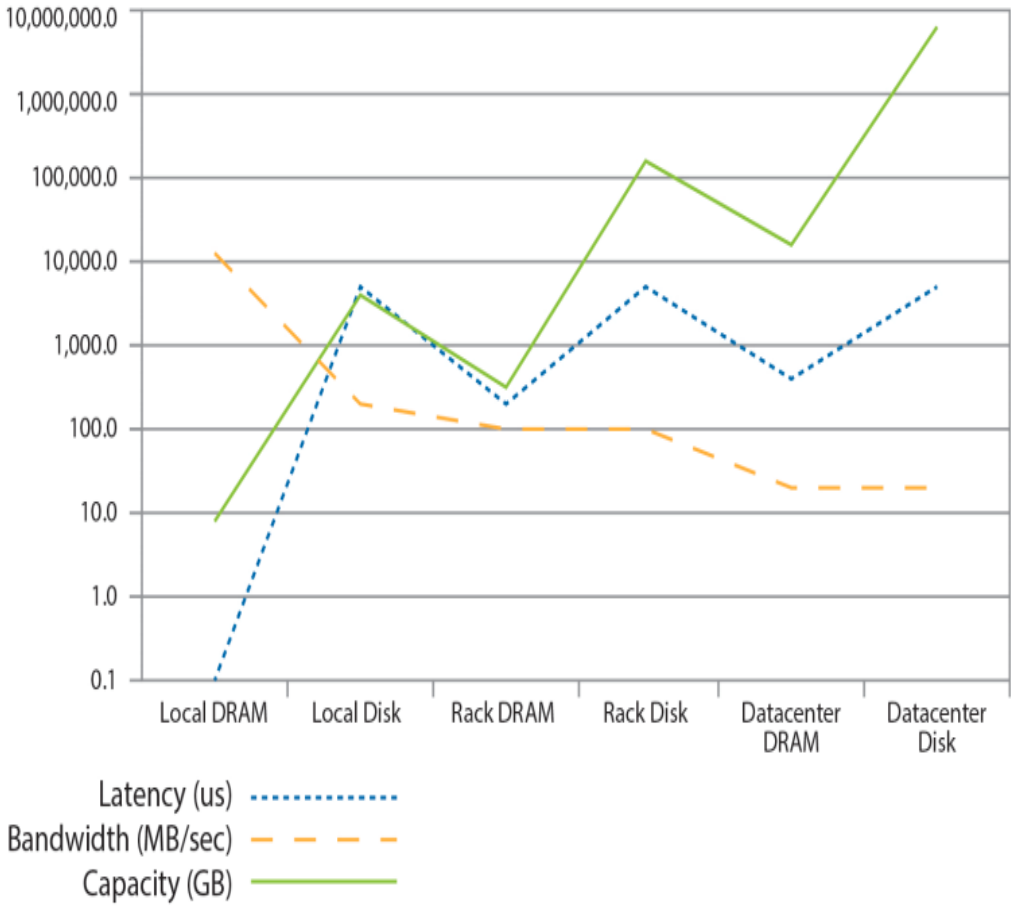
Why the latency of Rack DRAM is much higher than that of Local DRAM?



Answer: Networking.

Question 1b

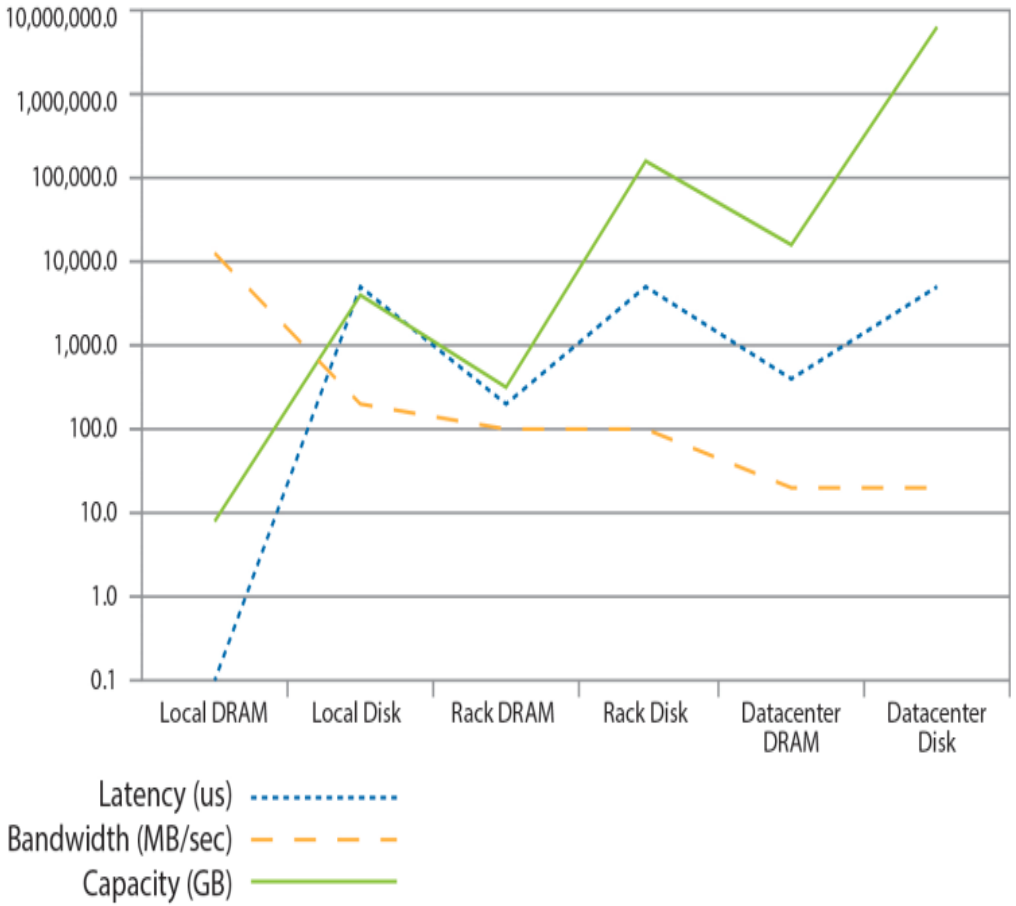
Why the latency of Rack Disk is similar to that of local Disk?



Answer: The network in the Rack is relatively fast, and the disk performance becomes the bottleneck.

Question 1c

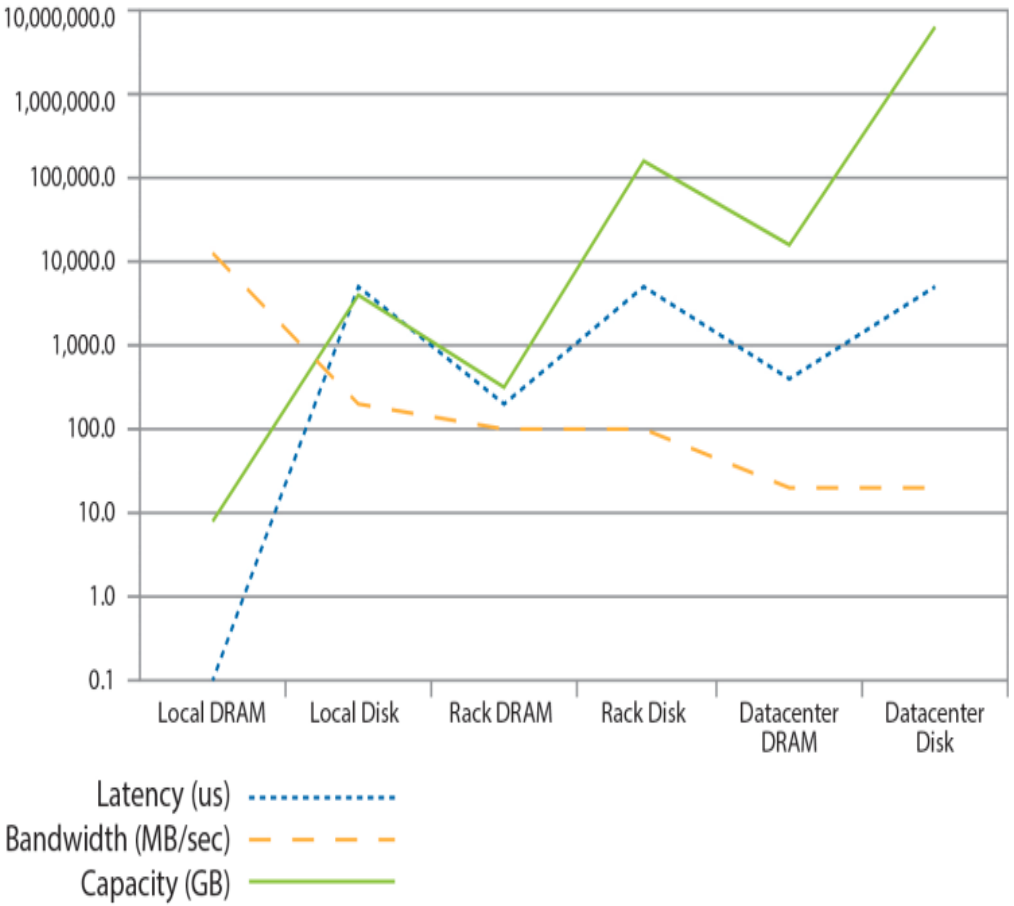
If we have an application with the working set smaller than Local DRAM, should we always put the working set into a single machine for efficiency



Answer: No. Depending on the gain of performing parallel computation on multiple machines.

Question 1d

Should we always put the data into DRAM?



Answer: No. DRAM is volatile (the data is off when power off), and disk can be used for persistency, and reliability execution.

Question 2

2. "A picture is worth a thousand words". An important aspect of big data is the presentation of the answers. In this question, we examine a number of plots from https://data.gov.sg/, which is an effort of public data analytics in Singapore. For each plot, write down your findings, share/discuss your findings with peer classmates and try to dig out more insights from discussion.

a) Crude birth rate

b) GDP at current market prices

c) Graduates from university first degree courses

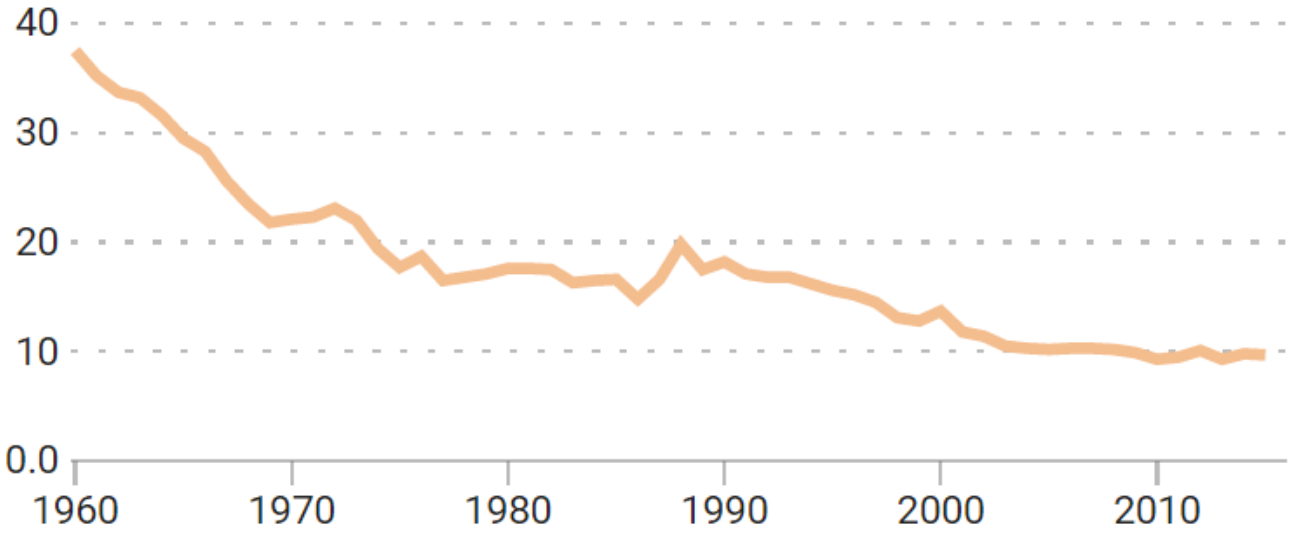
d) Rainfall - monthly total

Crude Birth Rate - Data

NUS National University of Singapore

9.7

Per Thousand Population in 2015



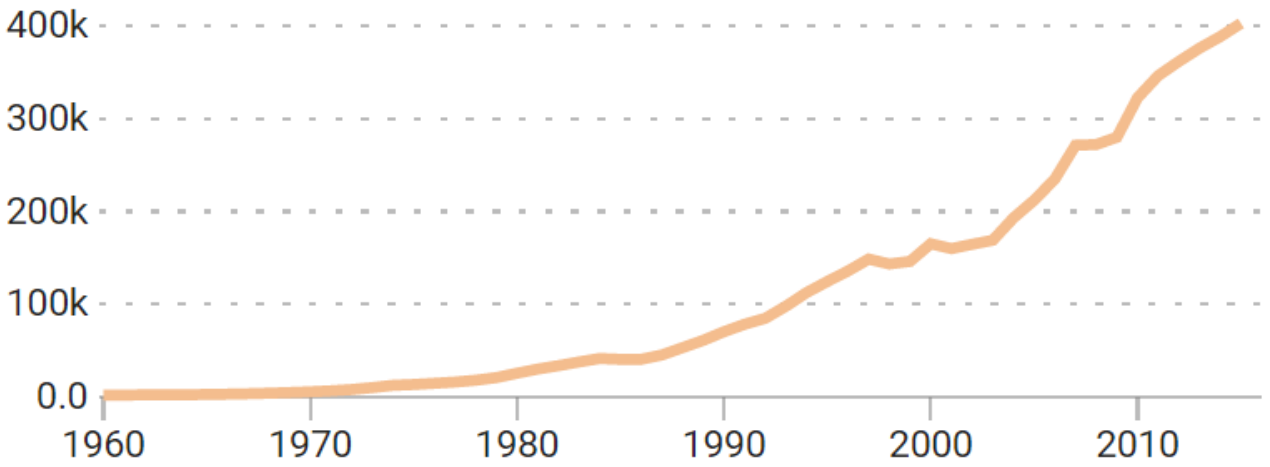
Answer: The birth rate has been going down since 1960. Now it is becoming very stably low.

GDP At Current Market Prices, Annual - Data

NUS National University of Singapore

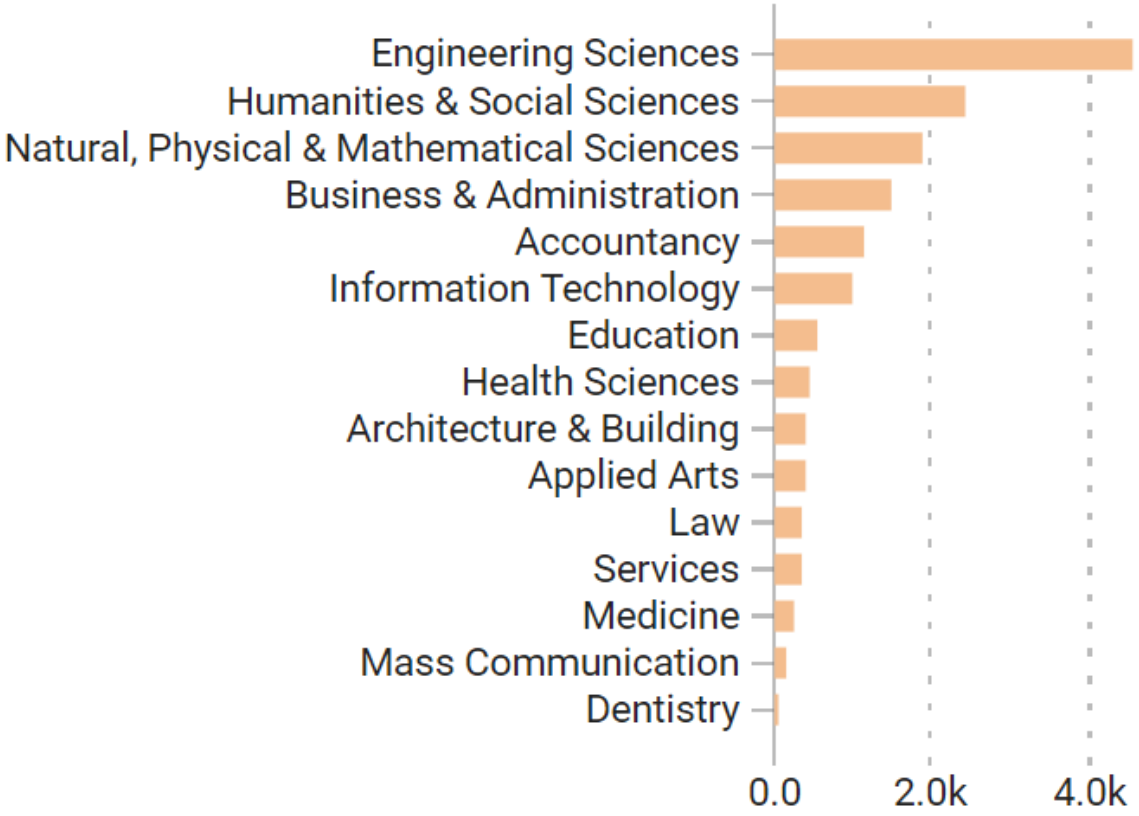
S$402,457.9

Million in 2015

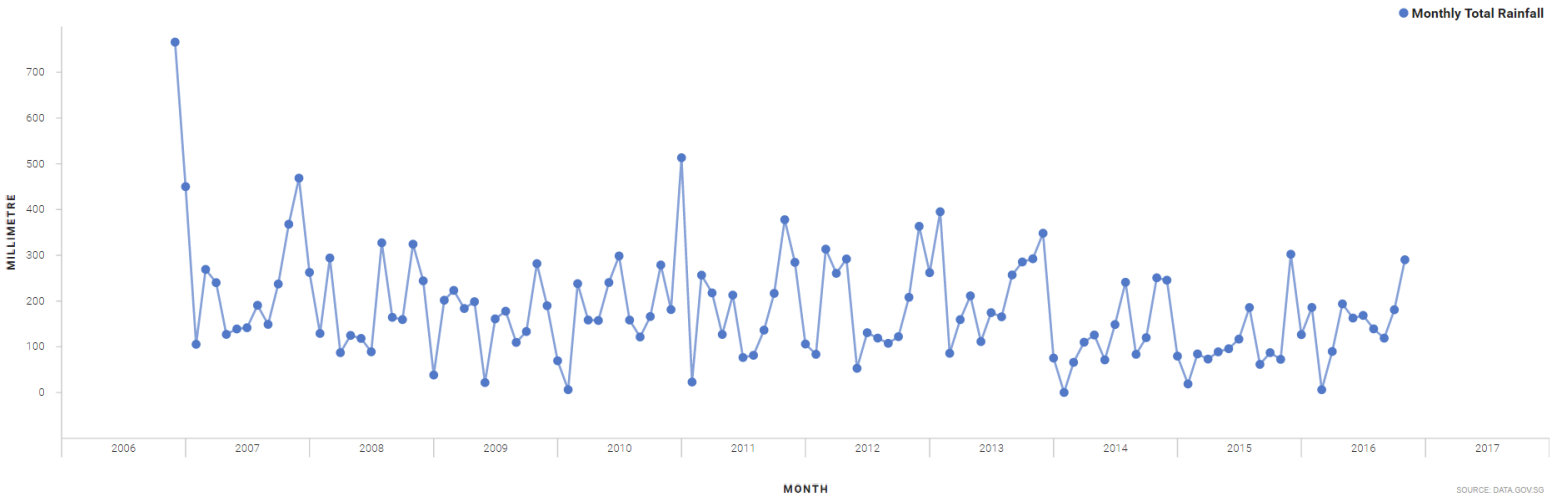


Answer: The GDP has been growing really fast. The latter years grows even at a higher pace.

Graduates From University First Degree Courses (2014) - Data



Answer: Long tails. The head is on engineering.



Answer: Within a year, the monthly rainfall fluctuates dramatically. Across years, the monthly rainfall also fluctuates quite significantly on the same month.

More Exercise Questions

Index	Question
1	Which challenges are important for gene sequencing in precision medicine? a) Volume b) Velocity c) Veracity d) Variety
2	When the number of tasks in a job is small, we should move the data to the task, rather than moving task to the data. a) True b) False
3	The major advantage of MapReduce is on its scalability and programmability. But, we may come out with non-optimal solutions for an application by using MapReduce programming. a) True b) False

Answer

Index	Question
1	Which challenges are important for gene sequencing in precision medicine? a) Volume b) Velocity c) Veracity d) Variety
2	When the number of tasks in a job is small, we should move the data to the task, rather than moving task to the data. a) True b) False
3	The major advantage of MapReduce is on its scalability and programmability. But, we may come out with non-optimal solutions for an application by using MapReduce programming. a) True b) False

Latency Numbers Every Programmer Should Know

https://people.eecs.berkeley.edu/~rcs/research/interactive_latency.html

Acknowledgement

Thanks to Li Qinbin for making these slides.

liqinbin@u.nus.edu