DLI Accelerated Data Science Teaching Kit

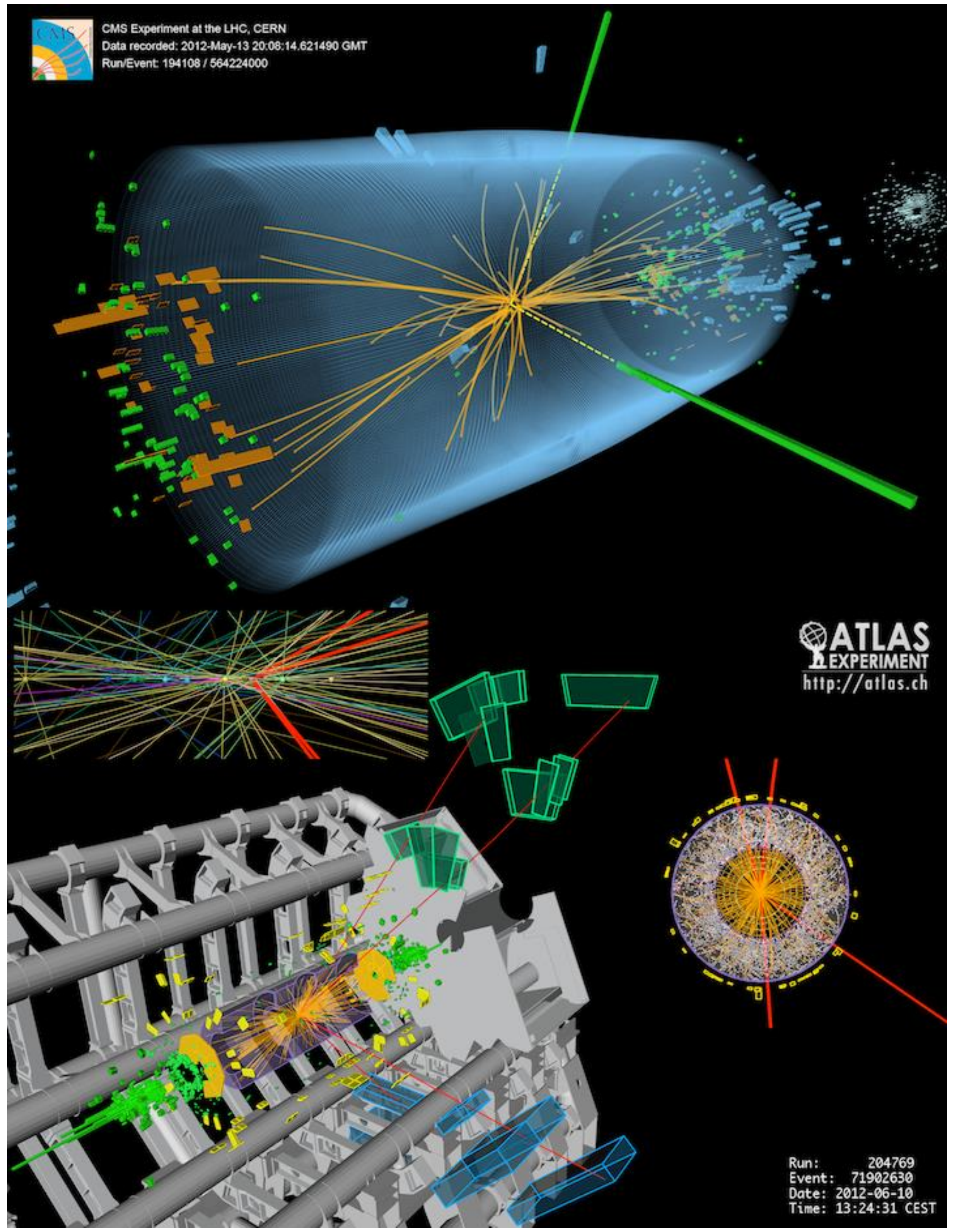
Lecture 10.1 - Big Data is Common. How to Store It?

How to Handle Data that is Really Big?

- Really big, as in...
  - Petabytes (PB, about 1000 times of terabytes)
  - Or beyond: exabyte, zettabyte, etc.
- Do we really need to deal with such scale?
  - Yes!

“Big Data” is Common...

- Google processed 24 PB / day (2009)
- Facebook’s add 0.5 PB / day to its data warehouses
- CERN generated 200 PB of data from “Higgs boson” experiments
- Avatar’s 3D effects took 1 PB to store
- So, think BIG!


CMS Experiment at the LHC, CERN
Data recorded: 2012-May-13 20:08:14.621490 GMT
Run/Event: 194108 / 564224000
ATLAS EXPERIMENT
http://atlas.ch
Run: 204769
Event: 71902630
Date: 2012-06-10
Time: 13:24:31 CEST

https://en.wikipedia.org/wiki/File:Avatar-Teaser-Poster.jpg

http://www.theregister.co.uk/2012/11/09/facebook_open_sources_corona/

http://thenextweb.com/2010/01/01/avatar-takes-1-petabyte-storage-space-equivalent-32-year-long-mp3/

http://dl.acm.org/citation.cfm?doid=1327452.1327492

How to Store Large Datasets?

First thing, how to **store** them?

Single machine? 60TB SSD announced. $$$$$…

Cluster of machines?

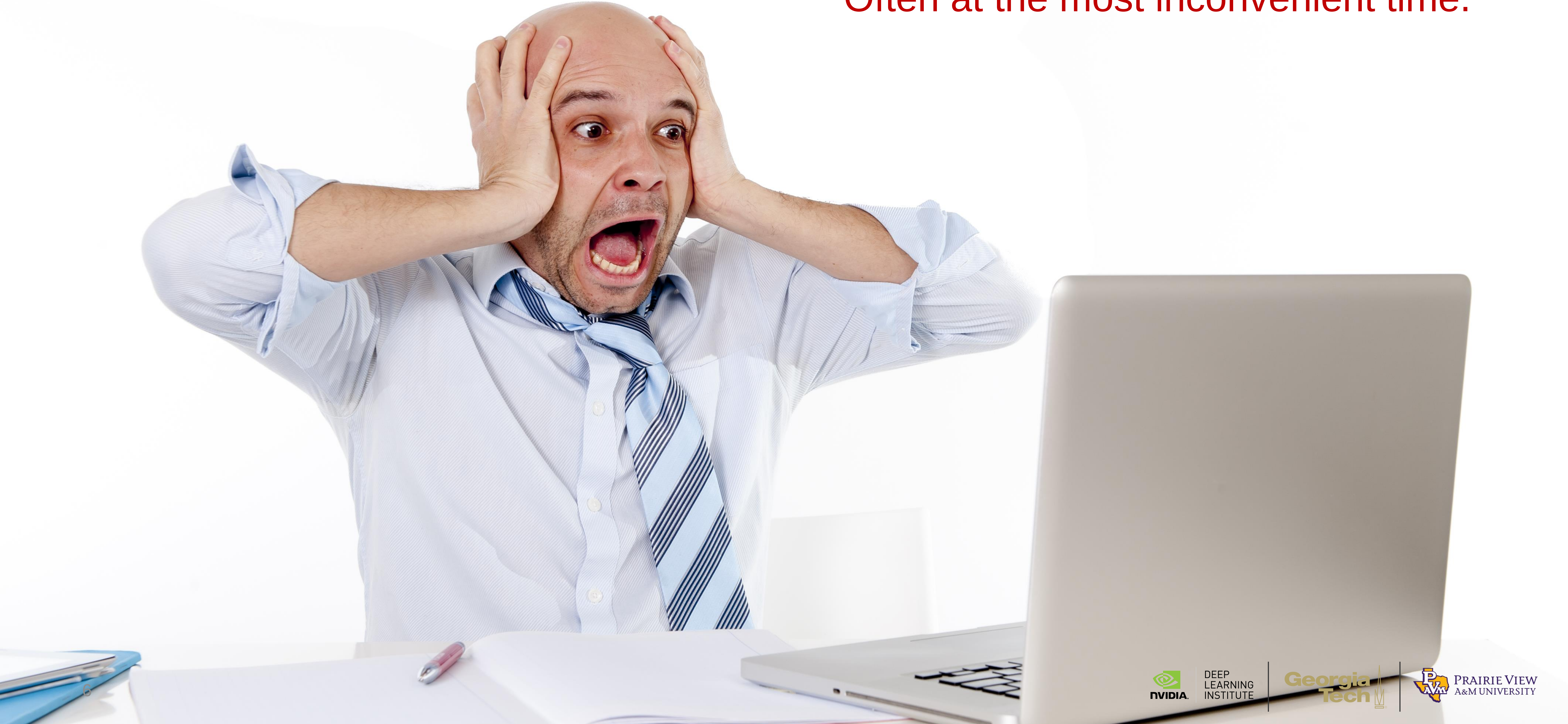
- How many machines?
- Need data backup, redundancy, recovery, etc.
- Need to worry about machine and drive failure.

Really? Really???

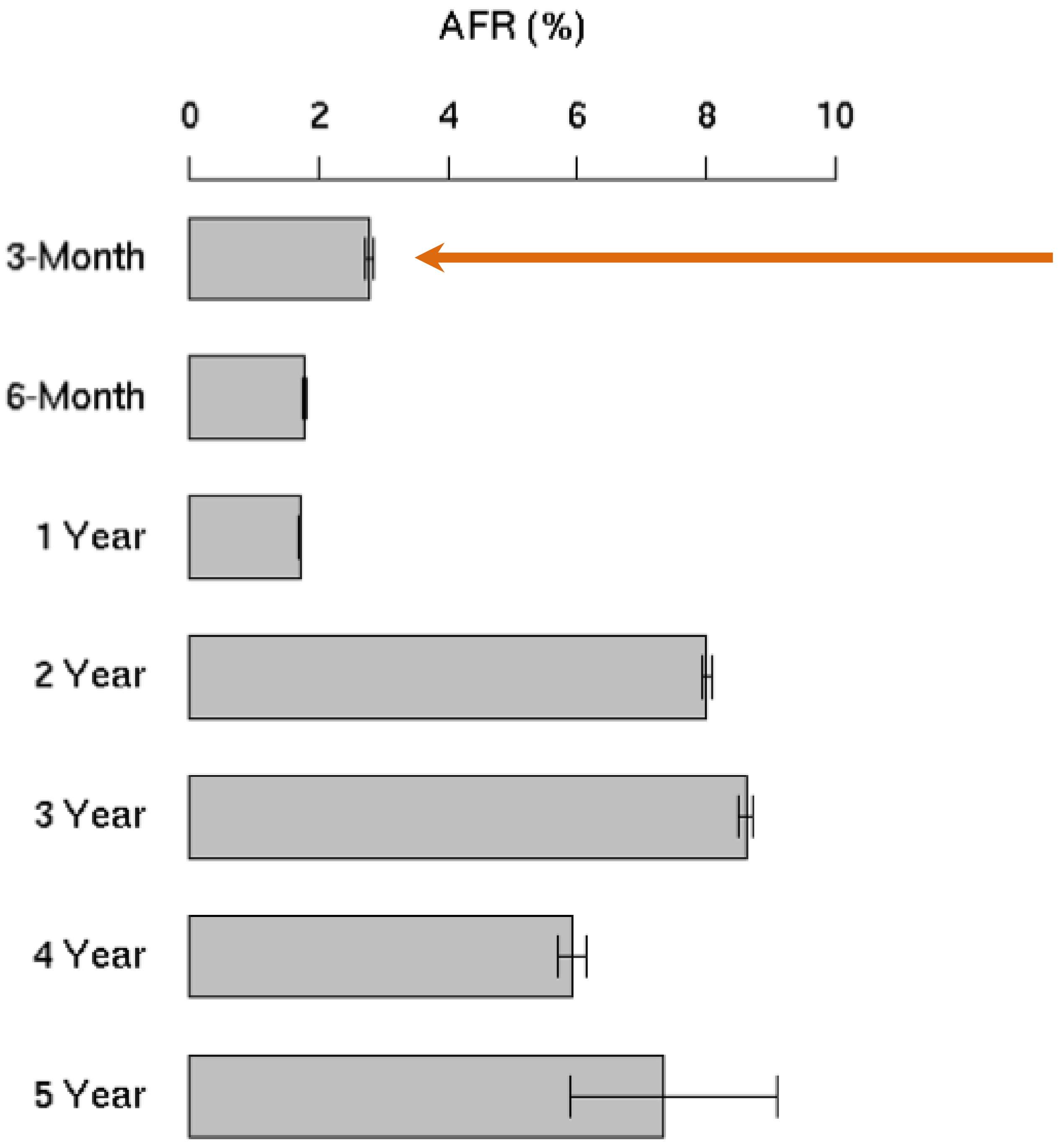
https://arstechnica.com/gadgets/2016/08/seagate-unveils-60tb-ssd-the-worlds-largest-hard-drive/

Computers and Disks Die!

Often at the most inconvenient time.



How Often Do Disks Fail?



3% of 100,000 hard drives fail within first 3 months

Figure 2: Annualized failure rates broken down by age groups

https://research.google.com/pubs/pub32774.html

NVIDIA. DEEP LEARNING INSTITUTE | Georgia Tech | PRAIRIE VIEW A&M UNIVERSITY

DLI Accelerated Data Science Teaching Kit

Thank You