Data Management – Nuts and Bolts

Don Johnson
Scientific Computing and Visualization

Overview

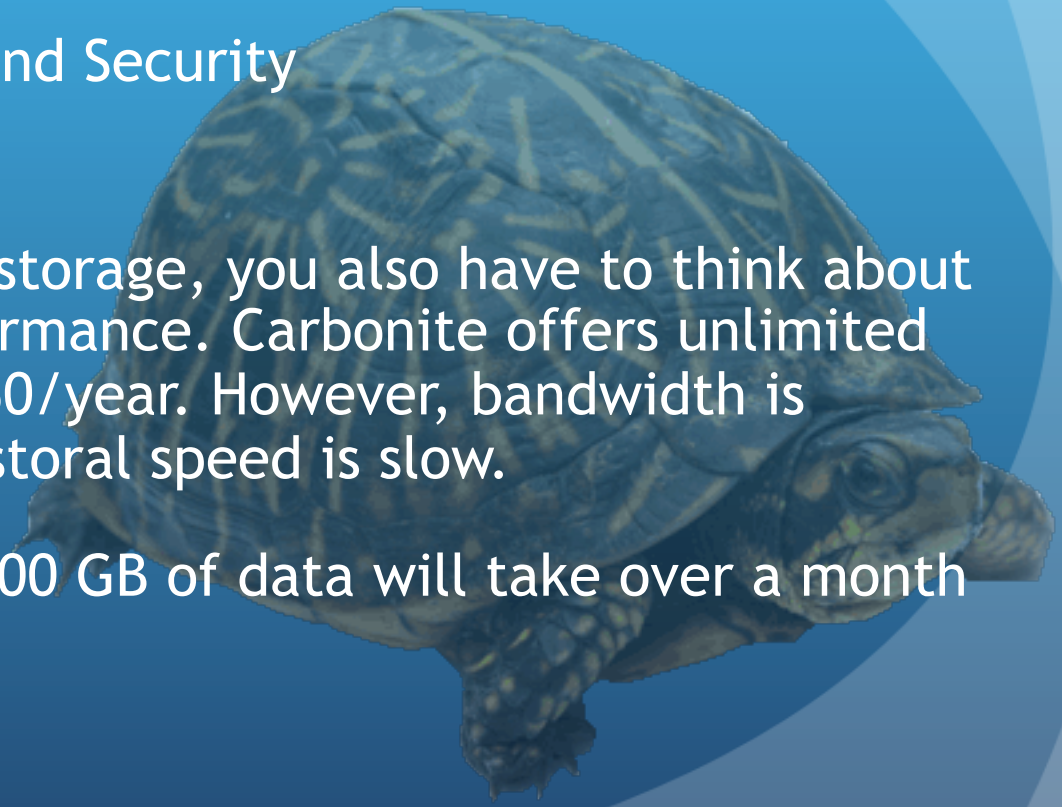
- Data Management
  - Storing data
  - Sharing data
  - Moving data
  - Tracking data (Client responsibility)
- Where can you obtain storage?
  - Retail
  - Online services – “The Cloud”
  - IS&T, College or Department

Attributes of Storage

- Capacity
- Performance
- Reliability, Safety and Security
- Cost

When thinking about storage, you also have to think about networking and performance. Carbonite offers unlimited backup storage for $60/year. However, bandwidth is throttled, and the restoral speed is slow.

At 14 GB/day, your 500 GB of data will take over a month to restore!



Attributes of Storage

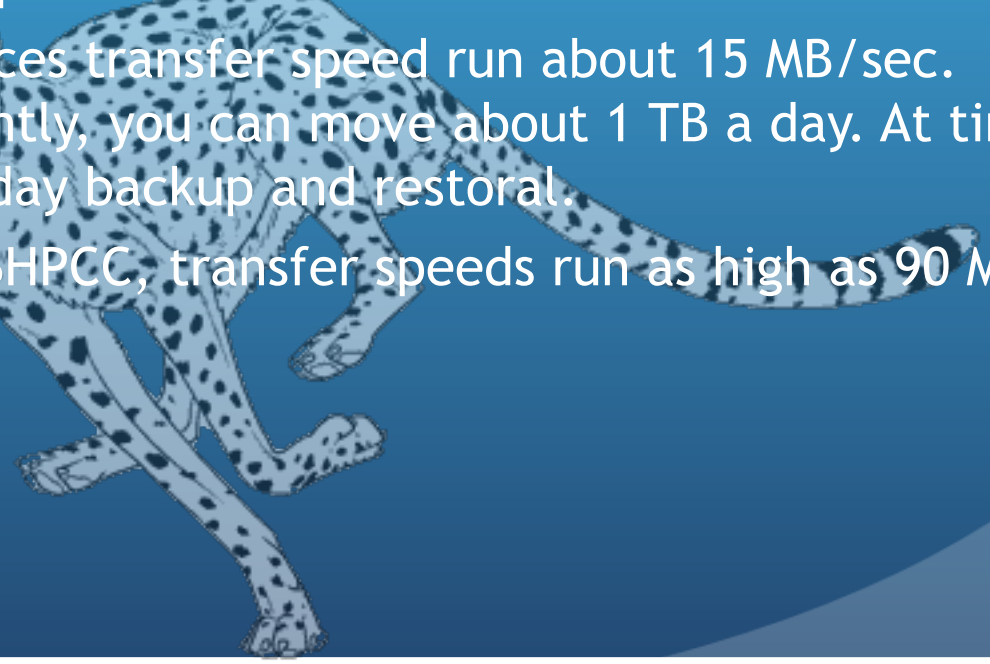
- Capacity
  - Sample Sizes of storage and stored items
    - *Word document or Excel Spreadsheet 500 KB*
    - *MP3 Song 5 MB*
    - *LANDSAT 8 170 km x 184 km scene 2 GB*
    - *MPEG2 Video 3 GB*
    - *Data stored in the human genome (summed across all cells) 150 trillion GB or 150 ZettaBytes*
  - The College of Arts and Sciences has 120 TB of public storage and will do one-on-one consulting for unique storage needs
  - Size of the MGHPCC storage 1+ PB

Attributes of Storage

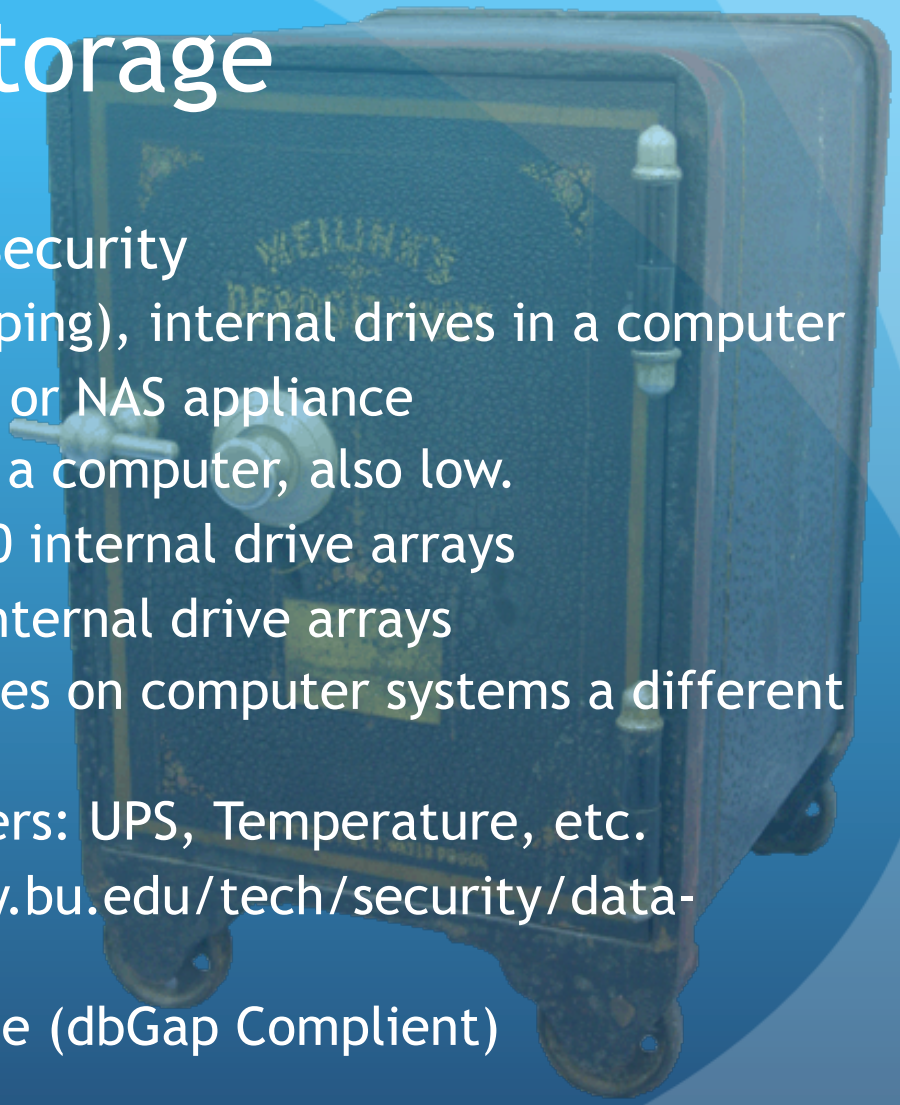
- Measuring Capacity - Units
  - Byte – 8 bits of information
  - KiloBytes (KB) 1x10^3
  - MegaBytes (MB) 1x10^6
  - GigaBytes (GB) 1x10^9
  - TeraBytes (TB) 1x10^12
  - PetaByte (PB) 1x10^15
  - ExaByte (EB) 1x10^18
  - ZettaByte (ZB) 1x10^21

Attributes of Storage

- Performance
  - Storage uses Bytes and Networking uses Bits
  - Comcast Internet service offers 4 Mb/sec up and 15 Mb/sec down. 100 Mb/sec equals 12.5 MB/sec, consequently, 15 Mb/sec equals less than 2 MB/sec.
  - In BU Offices transfer speed run about 15 MB/sec. Consequently, you can move about 1 TB a day. At times we get 2 TB/day backup and restoral.
  - On the MBHPCC, transfer speeds run as high as 90 MB/sec!



Attributes of Storage



- Reliability, Safety and Security
  - Very Low: RAID 0 (stripping), internal drives in a computer
  - Low: Any USB attached or NAS appliance
  - JBOD internal drives in a computer, also low.
  - Moderate: RAID 1, 5, 10 internal drive arrays
  - High: RAID 1, 6, ZFS2 internal drive arrays
  - BEST: Two or more copies on computer systems a different locations, mirrored
  - The environment matters: UPS, Temperature, etc.
  - Encryption http://www.bu.edu/tech/security/data-protection/
  - Restricted Project Space (dbGap Complient)

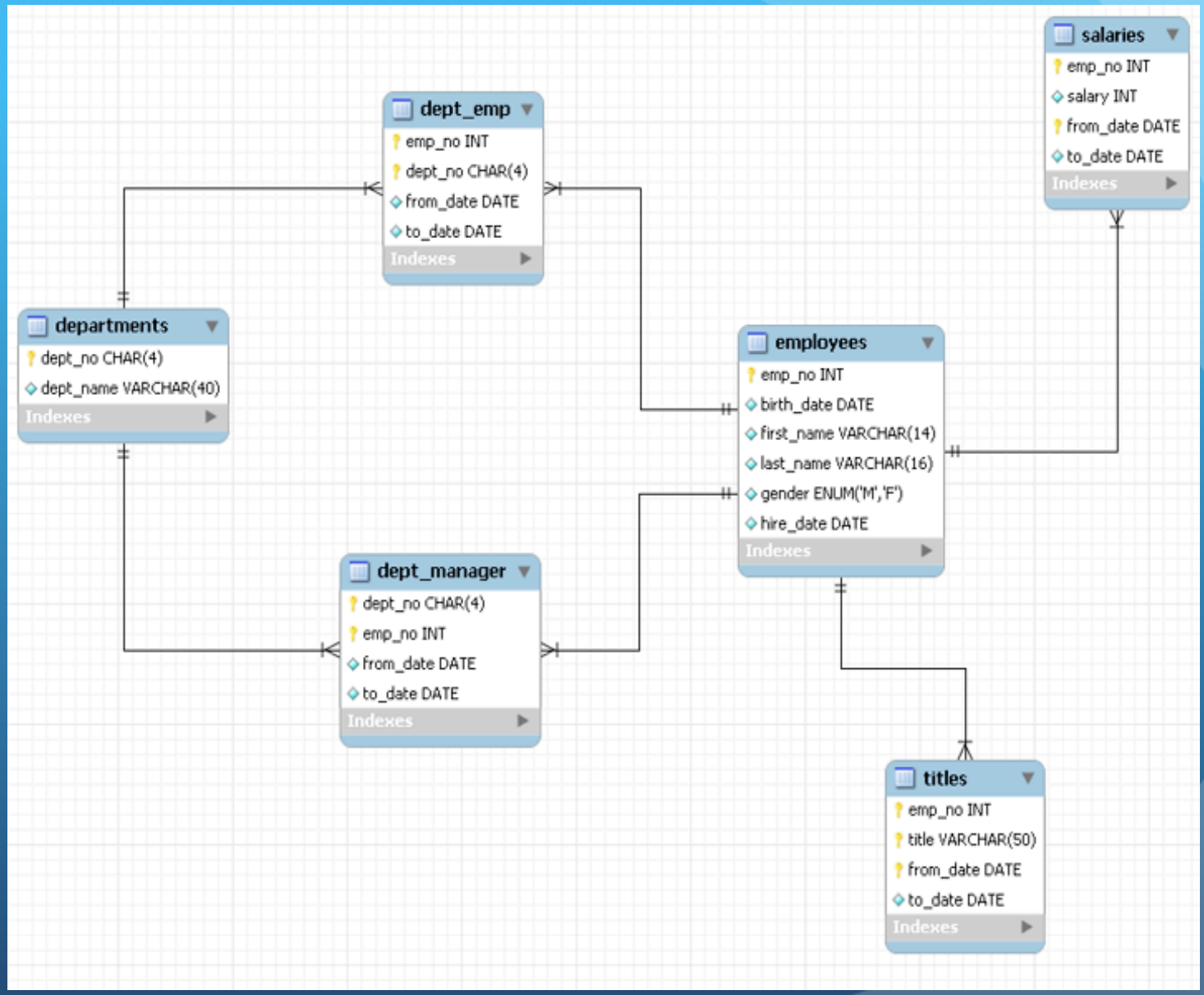
Attributes of Storage

- Cost
  - Rent
    - 1 TB free per project, Principal Investigator applies
    - ~$170 per 1 TB/year for project disk space, non-backed up, but hardware redundant, storage if IDC charge is included
    - 1 TB free per researcher of IS&T Archive Service storage – low performance, backup purposes
    - Possible free College provided storage for researchers without grant money to pay for storage
    - ~$1000 TB/year for virtual machine attached storage
  - Buy
    - ~$50 per TB for Buy-in model storage. Storage must be retired after five years. Only purchasable during Buy-in cycle
    - Seagate Backup Plus 4 TB USB $160 or ~$40/TB client attached storage
    - WD My Cloud 3 TB NAS: Network Attached Storage. These units are independent. $180 or ~$60/TB

Types of Data

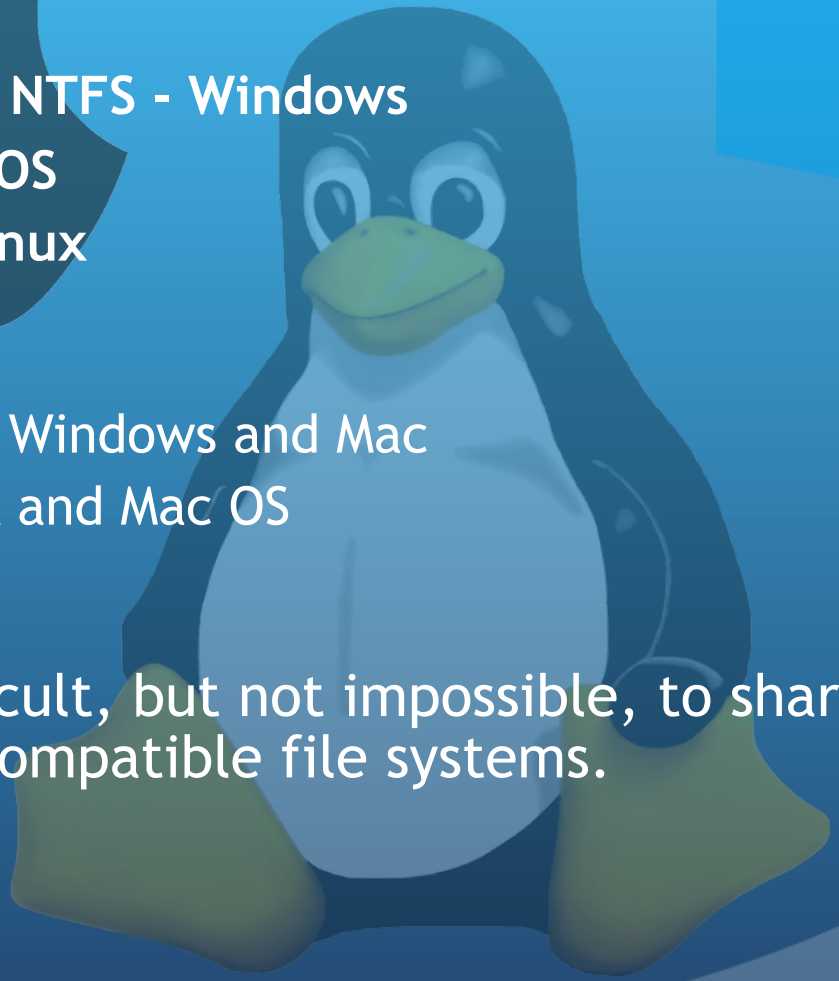
- Unstructured and Semi-structured
- Structured
  - Using database vs writing unstructured files
  - Tables (entities), Records (rows) and Fields (columns)
  - Relational vs NoSQL
    - *Relational database set up relationship between entities. Ex: Customer, Item, Order*
      - *MySQL http://www.mysql.com/*
      - *PostgreSQL http://www.postgresql.org/*
    - NoSQL
      - MongoDB http://www.mongodb.org/
- Versioned Data – Git and GitHub, Subversion

Structured Data - Database



File Systems and Sharing Data

- Local
  - **FAT32 and NTFS - Windows**
  - **HFS – Mac OS**
  - **Ext3,4 - Linux**
- Network
  - SMB/CIFS – Windows and Mac
  - NFS – Linux and Mac OS



It is more difficult, but not impossible, to share data residing on incompatible file systems.

Data Management Tools

- Service Request
- Requesting Project Storage
- Navigating the Cluster File System
- FTP, SFTP: Fetch and WinSCP
- Connecting to a PostgreSQL or MySQL server
- Evernote
- Mekentosj Papers
- Git and GitHub

Service Requests

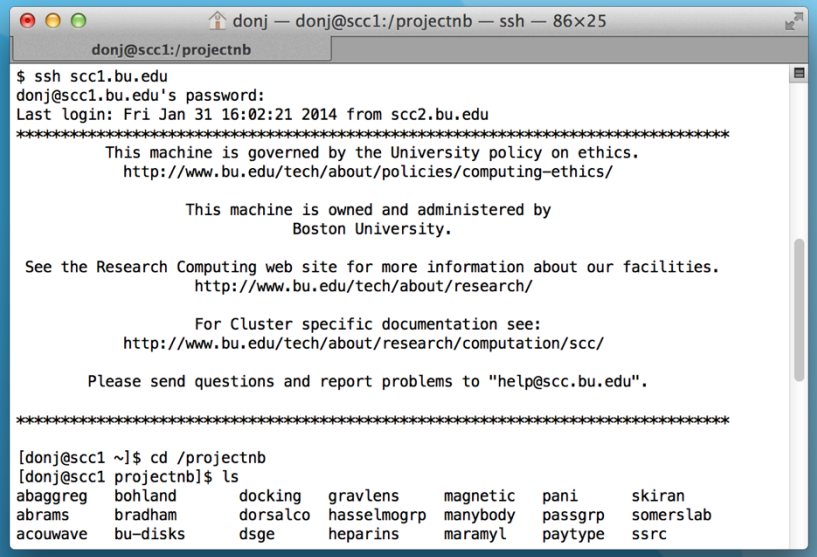
- Request space on FTP server to share data
- If you are “mounting” the space locally on your desktop, you must specify, “NFS” or “SMB”
- Go to http://www.bu.edu/tech/ then

  - Research Computing
  - General
  - “Tell us how we can help”
  - Click on “track this request and add information online’
  - Write down the incident number e.g. “INC11472245” so you can track status

Requesting Project Storage

- This is for Principal Investigators only. Doctoral students, post-docs and visiting scholars have to request storage or other services through their affiliated PI, or follow the procedure on the previous slide, “Service Request.”
- Decide whether to “Rent” or “Buy”
- Go to http://www.bu.edu/tech/accounts/special/research/applications/
- Or go to http://www.bu.edu/tech/about/research/computation/file-storage/ and search for the “Buy-in options” link in the “Project Disk Space” section

Navigating the Cluster File System



- Log in using a terminal program
- Your project data will be at:
  - /net/<server_name>
  - Or /projectnb/<project_name>
  - Or /project/<project_name>
- Use “cd”, “find”, “|” (pipe) and “grep” to locate your data
- Use “scp”, “rsync” and “mv” to move and copy your data
- Sign up for the the SCV Linux Class: http://www.bu.edu/tech/about/research/training/live-tutorials/

Connecting to a PostgreSQL or MySQL Server

- Via the command line using “psql” or “mysql” and then issuing SQL commands
- GUIs available
  - pgAdmin for PostgreSQL – runs locally, Windows or Mac OS clients available
  - http://www.phpmyadmin.net/home_page/index.php
  - phpMySQL for MySQL
    - Runs on the locally or a server
    - http://www.phpmyadmin.net/home_page/index.php

FTP and SFTP: A Way to Share and Move Data

- FTP allows anonymous log in
- SFTP is encrypted and secure
  - Access data on the cluster file system
  - Transfer data between systems and your desktop
  - Use GUI programs or via command line using “scp”
- Fetch for Mac OS is free via the BU website: http://www.bu.edu/tech/support/desktop/distribution/ftp/
- FileZilla for Windows is Open Source software: http://www.bu.edu/tech/support/desktop/software/windows/filezilla/

- Store unstructured data in the cloud: sound, video, images, hand written notes, web pages
- Web interface or install Mac OS, Windows, iOS and Android clients having extra functionality
- Free for most usage
- https://evernote.com/

Mekentosj Papers

- Organizes journal articles and PDF files
- Download journal articles from BU Libraries directly into the application
- http://www.papersapp.com/

- Store versioned data in the cloud: source code, configuration files and documents
- Use command line, built-in functionality in applications, or web interface
- Free for public repositories, $7/month for five private repositories
- https://github.com/
- Take the IS&T Tech Lunch tutorial being offered next month on Mar 18 from 11:00 a.m. to 3:00 p.m.

Final Questions?