CS 199 ACC
Data

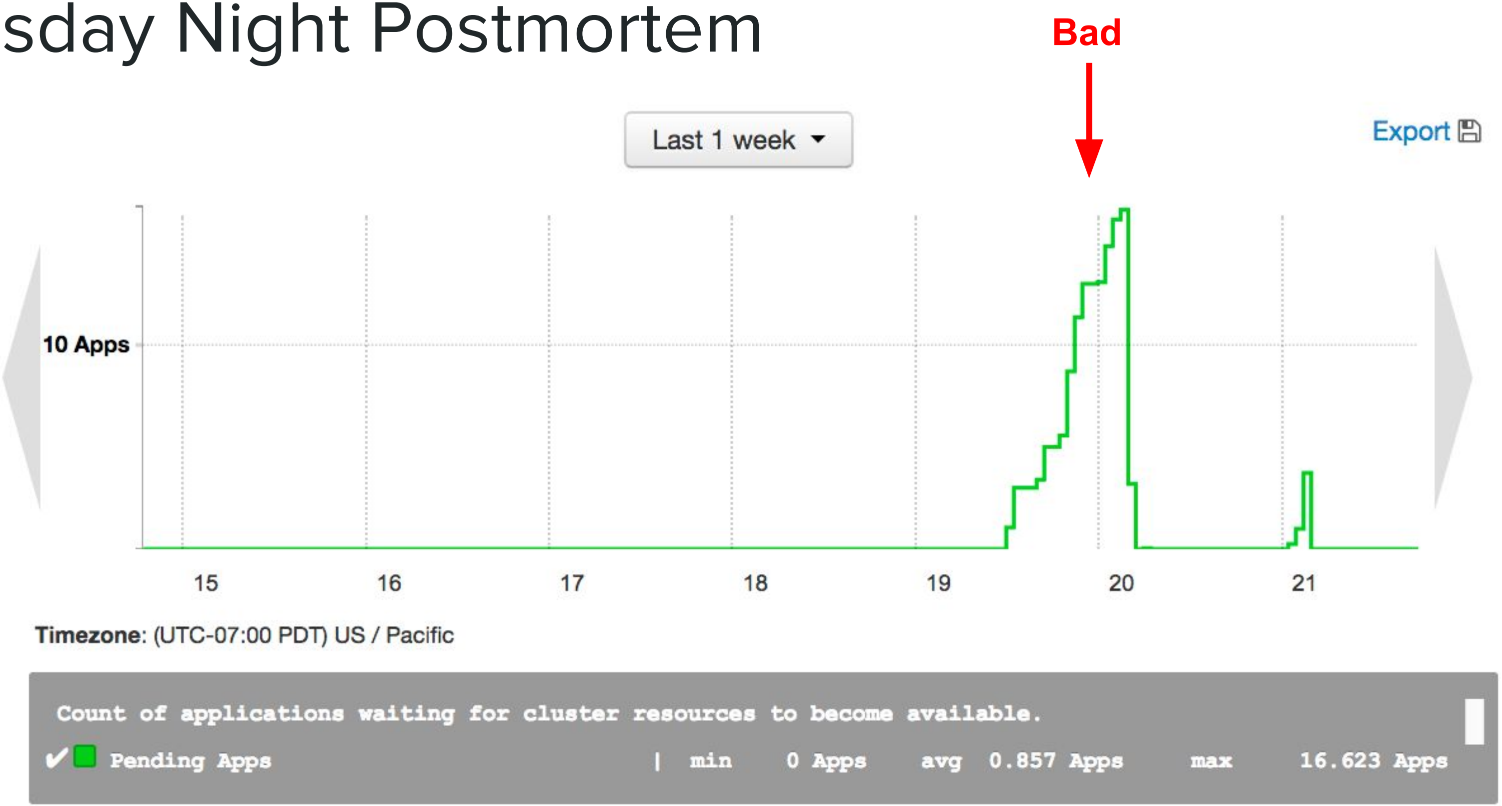
Prof. Robert J. Brunner
Sameet Sapra
Ben Congdon
Tyler Kim
Bhuvan Venkatesh

MP 2

How was/is it?

Cluster Postmortem

Tuesday Night Postmortem

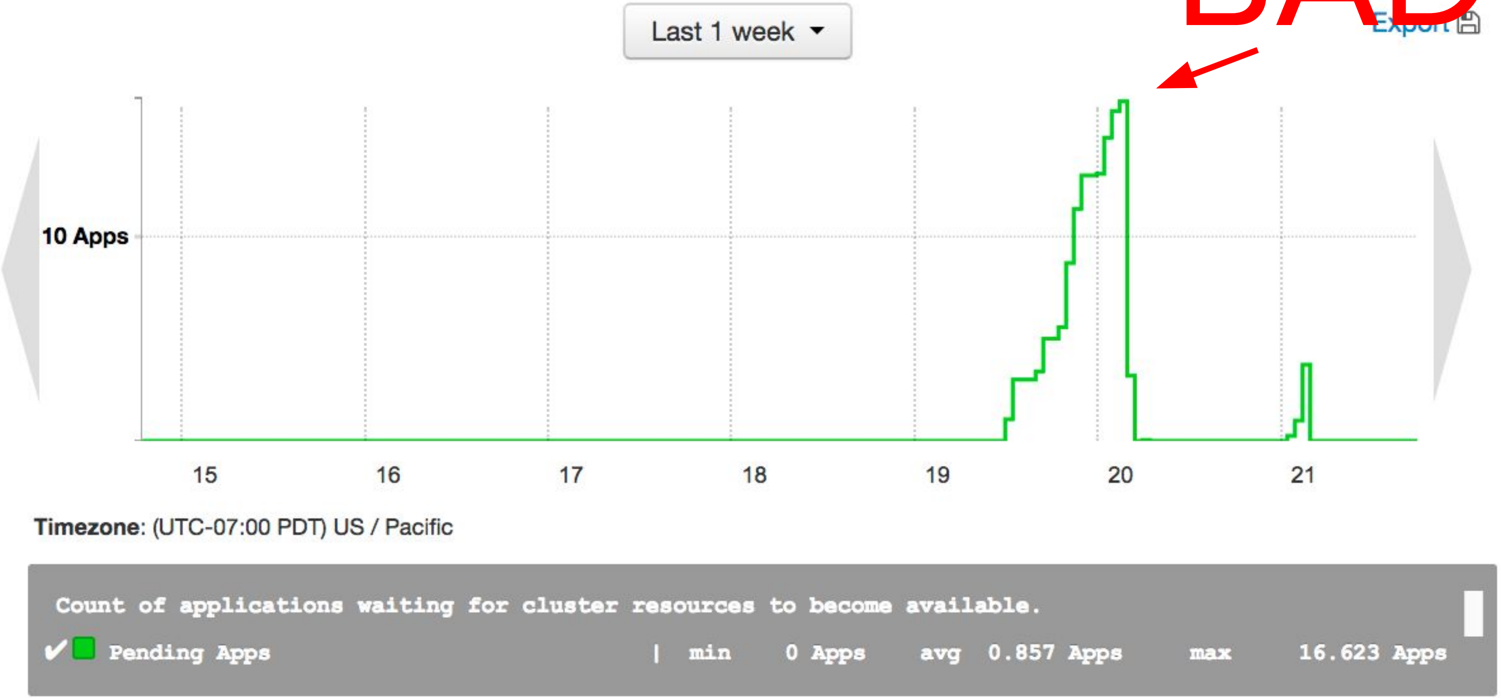


Bad

Timezone: (UTC-07:00 PDT) US / Pacific

```
Count of applications waiting for cluster resources to become available.
✔🟩 Pending Apps                  |  min    0 Apps    avg  0.857 Apps    max    16.623 Apps
```

Tuesday Night Postmortem



BAD

Timezone: (UTC-07:00 PDT) US / Pacific

```
Count of applications waiting for cluster resources to become available.
✔🟩 Pending Apps                    |  min    0 Apps   avg  0.857 Apps    max    16.623 Apps
```

Tuesday Night Postmortem

- **Root cause**: Really slow job hogged cluster resources
- **Secondary Cause**: Jobs were allowed to request too many resources
- **Secondary Cause**: Jobs were not set to timeout
- **Tertiary Cause**: Container sizes were too large

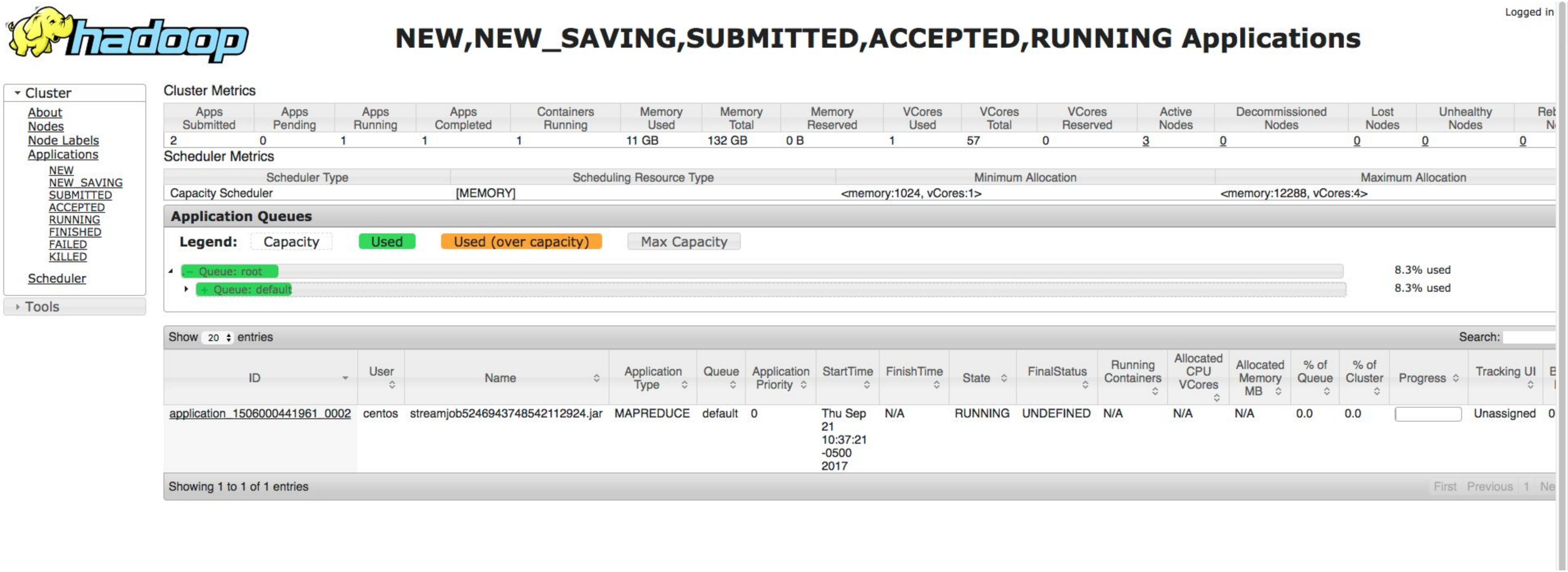
- Cluster Issues on Thursday Morning: Mistakenly set task timeouts too low
  - 300 milliseconds vs. 300 seconds. Whoops, our mistake...

New Policies

- Jobs will be killed after 30 minutes of runtime
- You can only have 1 job running at a given time

Accessing the Web Interface

- Use SSH Tunneling to access the internal web interface ports
- Tutorial on the MP2 docs



Logged in

NEW,NEW_SAVING,SUBMITTED,ACCEPTED,RUNNING Applications

Cluster Metrics

Apps Submitted	Apps Pending	Apps Running	Apps Completed	Containers Running	Memory Used	Memory Total	Memory Reserved	VCores Used	VCores Total	VCores Reserved	Active Nodes	Decommissioned Nodes	Lost Nodes	Unhealthy Nodes	Reb N
2	0	1	1	1	11 GB	132 GB	0 B	1	57	0	3	0	0	0	0

Scheduler Metrics

Scheduler Type	Scheduling Resource Type	Minimum Allocation	Maximum Allocation
Capacity Scheduler	[MEMORY]	<memory:1024, vCores:1>	<memory:12288, vCores:4>

Application Queues

Legend: Capacity Used Used (over capacity) Max Capacity

- Queue: root 8.3% used
  - + Queue: default 8.3% used

Show 20 entries Search:

ID	User	Name	Application Type	Queue	Application Priority	StartTime	FinishTime	State	FinalStatus	Running Containers	Allocated CPU VCores	Allocated Memory MB	% of Queue	% of Cluster	Progress	Tracking UI	B
application_1506000441961_0002	centos	streamjob5246943748542112924.jar	MAPREDUCE	default	0	Thu Sep 21 10:37:21 -0500 2017	N/A	RUNNING	UNDEFINED	N/A	N/A	N/A	0.0	0.0		Unassigned	0

Showing 1 to 1 of 1 entries

A little bit about being a good cluster neighbor

- Ideally, we would give everyone their own cluster, but we have to share the cluster
- So, to avoid problems like Wednesday night. We will try to give time estimates on each of the assignments. If your job runs longer than that, **you should kill it** (please).
- Before you log out, make sure you don’t have any stray jobs running by running `yarn top`. If you do, just kill them
- YARN won’t give us any notifications but please let us know if this happens

Killing Hadoop Jobs

- List Jobs:
  - hadoop job -list

- Killing Jobs:
  - hadoop job -kill <job_id>

Onto Focus - Data

This Week

- Data!
- Getting Data (Legally)
- Data Carwash (Cleaning Data)
- Using Data

Data Licensing

- Can you use any data?
- Check data sources for restrictions (commercial uses, foreign uses, etc)
- MIT License
  - You can do anything with it as long as you keep the license and copyright
  - One of the most easy to use licensing
- GPLv2/v3
  - If you use it, you must distribute the source of anything built with it
  - Must include copyright, license, link to the original, and details of your changes
  - Sorta like a virus, anything that uses it must then become GPL
    - Actually a good thing when you want to keep software/data completely open
    - But companies typically hate it for obvious reasons

Data Sources

- Data is everywhere! We can track things at scales that we never have before
- There are primarily two types of data
- Refined Datasets
    - These are typically academic or governmental datasets that are released to the public that have been pruned for a specific purpose.
    - You may need to reprune it for your purpose, but for the most part all the cases of missing values and parsing is done for you (the data is in a table-like format)
- Raw Data
    - This can be any data source: facebook, twitter, or reddit. You will definitely need to clean the data in order to get it to a state where it actually means something in aggregate

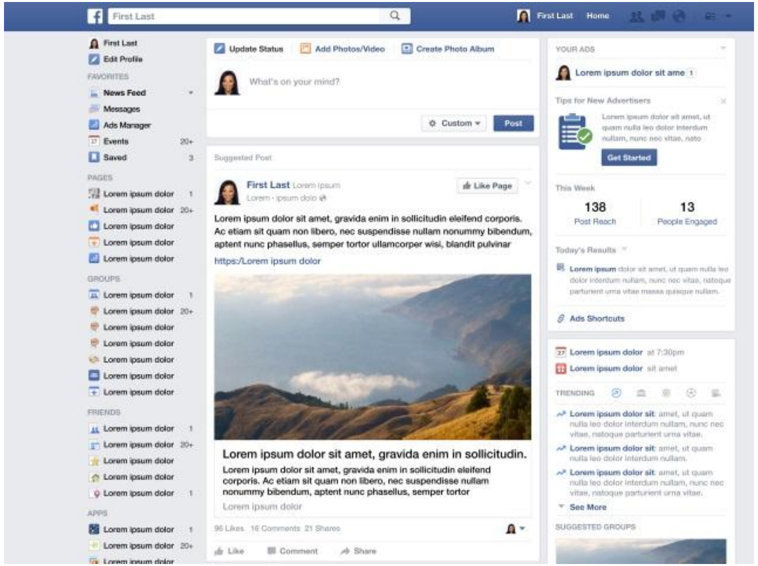
Premade Dataset

- There are a lot of good data sources already
- If you can, use them instead of getting your own data
  - It has been cleaned
  - It is typically easier to download
  - There are other papers you can look to for examples of how they manipulated it
- You may want to combine datasets/fill in NAs but for the most part they are well behaved.

https://github.com/cazala/mnist

Raw Data

How would you process your newsfeed?



First Last Lorem ipsum
Lorem · ipsum dolo

Lorem ipsum dolor sit amet, gravida enim in sollicitudin eleifend corporis. Ac etiam sit quam non libero, nec suspendisse nullam nonummy bibendum, aptent nunc phasellus, semper tortor ullamcorper wisi, blandit pulvinar

https:/Lorem ipsum dolor

Lorem ipsum dolor sit amet, gravida enim in sollicitudin.

Lorem ipsum dolor sit amet, gravida enim in sollicitudin eleifend corporis. Ac etiam sit quam non libero, nec suspendisse nullam nonummy bibendum, aptent nunc phasellus, semper tortor

Lorem ipsum dolor

96 Likes 16 Comments 21 Shares

Lorem ipsum dolor sit ame 1

Lorem ipsum dolor sit amet, ut quam nulla leo dolor interdum nullam, nunc nec vitae, nato

Lorem ipsum dolor sit amet, ut quam nulla leo dolor interdum nullam, nunc nec vitae, natoque parturient urna vitae massa quisque nullam.

Lorem ipsum dolor at 7:30pm

Lorem ipsum dolor sit amet

Lorem ipsum dolor sit amet, ut quam nulla leo dolor interdum nullam, nunc nec vitae, natoque parturient urna vitae.

Lorem ipsum dolor sit amet, ut quam nulla leo dolor interdum nullam.

Lorem ipsum dolor sit amet, ut quam nulla leo dolor interdum nullam, nunc nec vitae, natoque parturient urna vitae.

http://static.adweek.com/adweek.com-prod/wp-content/uploads/sites/2/2016/02/NewsFeedTeaser640.jpg

Raw Data - Web Scraping

- Ideally you start with a headless browser and download a subset of the javascript objects/html from the page
- Once you have the objects, save them in some format that makes sense. This can be a CSV, a Table, what have you.
- You may need to do some HTML parsing in this case. Get yourself an HTML parser and write all the relevant files

20 Second Example

```python
from bs4 import BeautifulSoup
soup = BeautifulSoup("""<html><body><ul>
    <li class="shoe-item">Air Jordans</li>
    <li class="shoe-item">Light up Sketchers</li>
    </ul><body></html>""", 'html.parser')
for item in soup.find_all('.shoe-item'):
    print(item.inner_html)
```

Scraping Problems

- You may get rate limited
- You may get blocked
- You may be violating the law
- Whenever the HTML changes, your code immediately breaks

http://4.bp.blogspot.com/--POGMPcheZQ/U3AFerYzrKI/AAAAAAAABDo/uNPZGsCpGAI/s1600/web+scraping+jime.jpg

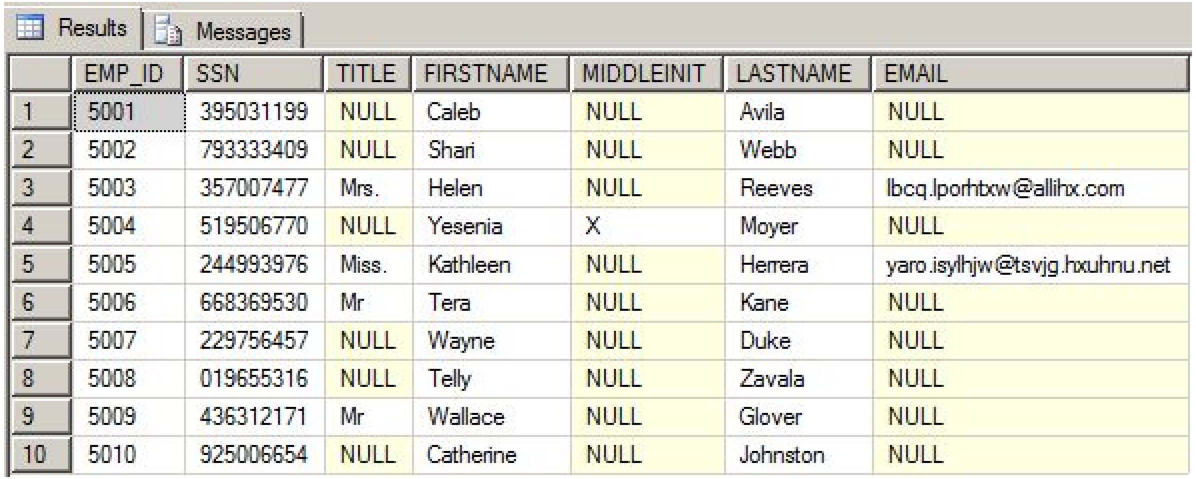
Data Carwash

Why do we need to clean the data?

- Data could have
  - Missing Values
  - Duplicate Values
  - Invalid Values
  - Useless Values
  - Etc Etc
- We need to make sure that the data that we give to the machine learning algorithm is as close to representative as possible

For the purposes of this lecture

- We want the data in TABLE format, which looks like this



	EMP_ID	SSN	TITLE	FIRSTNAME	MIDDLEINIT	LASTNAME	EMAIL
1	5001	395031199	NULL	Caleb	NULL	Avila	NULL
2	5002	793333409	NULL	Shari	NULL	Webb	NULL
3	5003	357007477	Mrs.	Helen	NULL	Reeves	lbcq.lporhtxw@allihx.com
4	5004	519506770	NULL	Yesenia	X	Moyer	NULL
5	5005	244993976	Miss.	Kathleen	NULL	Herrera	yaro.isylhjw@tsvjg.hxuhnu.net
6	5006	668369530	Mr	Tera	NULL	Kane	NULL
7	5007	229756457	NULL	Wayne	NULL	Duke	NULL
8	5008	019655316	NULL	Telly	NULL	Zavala	NULL
9	5009	436312171	Mr	Wallace	NULL	Glover	NULL
10	5010	925006654	NULL	Catherine	NULL	Johnston	NULL

https://www.red-gate.com/simple-talk/wp-content/uploads/imported/1519-image003.png

Missing Values

- What can we do?
- We can drop data
    - But that may skew our dataset, especially if we have a lot of missing values.
- We can make an educated guess of the values
    - Hard to do for categorical data
    - We can do a simple replacement with the mean for numerical data
    - We can also sample a probability distribution and fill in the values for that (randomly)
- Some algorithms don’t need all the values figured out
    - In that case, we leave as is because we want to put as little of our bias into the data

Duplicated Data

- Answer may seem simple, deleted the data
- But how do you define similar vs the same. The exact same comment content on reddit could mean something. Especially if we are doing frequency analysis.
- Can't just use raw file machine on a series of N files because the resulting algorithm is **ϴ(n*log(n)).** We can barely handle linear.
- What if the unmatched data is in the contents of the file? Raw dynamic programming algorithm give us a **O(n^2),** but we can get **ϴ(n*log(n))** by playing around with multi tries.

Invalid and Useless Values

- Really depends on your use case
- You really need to see what context the value is (age > 0)
- For useless values, ideally you keep them in the database because they may be useful later. If you are pretty sure that they aren’t going to be used anymore, you can just discard them in the data
- Again highly subjective

Combining Datasets

- If you have one data set that has (name, tweet timestamp) and another from equifax that has (name, ssn) you may want to join those two datasets to see if your tweet timestamp has any correlation with your SSN (joke)
- But either way, you would have to write a python job, spark job, etc etc so that for each record you find the appropriate record in the other dataset and end up writing out one row
- You may run into a **lot** of missing values if you don’t have an SSN for name so you may want to employ some of the methods earlier or just keep them separate.

Evaluating Your Data

Data Evaluation

- Besides just plugging into a machine learning model, you may want to understand how good a dataset is.
- You may test sparsity, how many values are filled out.
- You may also just run it through your algorithm (given that you have one) and just see what the output is supposed to be.
- Usually really hard, especially if you combine non-homogenous datasets
- Harder to find bias without another dataset

Final Destination (Part 1)

- Your data then just needs to be written to your platform for which you already have you analytics script ready.
- If you already have a hadoop MR script, then just copyFromLocal and run your mapreduce job.
- Same goes for any other platform like MYSQL, Spark, what have you
- Data may be big, so you may need to use a service like AWS data transfer or stream into your Hadoop/Cassandra cluster very slowly
- Or if you want to append to already existing data, do it in batches

Last Thing - (De)anonymization

- If you have Personally Identifiable Information (PII) in your dataset, ideally you want to censor that. If someone has a disease, instead of putting their name and disease, give a name_id and disease_id and make sure to document that those are categorical variables
- Hard problem because a malicious user could be able to deanonymize your data (ie the most common disease is the common cold and the next one is … and so on). Try to use the latests encryptions (AES) or secure hashes (SHA-2) and make sure there are no unwanted collisions, use a random salt.

MP 3

Due in one week (9/27) at 11:55pm

- We’re giving you a large set of tweets from Twitter over several months
- This is the first real dataset which you do not want to run on your laptop
- We want to stress test the server more. For this week’s lab only use the cluster for testing
- Use `head` or `tail` so you don’t test on all the data at once

Debugging Hadoop

The stack traces are very very long, but there are a few easy ones

Error Launching job: Input path does not exist:

Fix: Load the input file into HDFS using hdfs dfs -copyFromLocal

Error Launching job: Output directory hdfs://192-168-100-234.local:8020/out3 already exists

Fix: Delete the folder or rename your output folder

```
hdfs dfs -rm -r out3
```

More debugging

java.lang.RuntimeException: PipeMapRed.waitOutputThreads(): subprocess failed with code 1

- If it happened and map % was less than 100% it's an error in your map script
- If it happened and map % == 100 it's probably your reducer script

Use unix commands to test on a sample of data

- head book.txt | ./mapper.py # for mapper error
- head book.txt | ./mapper.py | sort -n -k 1 | ./reducer.py | sort -n -k 2 | tail -10 # for reducer error

More debugging

- Rule of Thumb: It’s usually only a cluster error if you have issues submitting your job.
- Once the job is Running, it’s *probably* your code that’s causing an issue
- If you’re job is taking a long time to start (but isn’t failing), check the job queue. It’s likely that the cluster is out of additional resources and you will need for another student’s job to complete