NYU | Center for Data Science

Week 01.1: Introduction

DS-GA 1004: Big Data

Instructor: Brian McFee

The ~~three~~ four “V”s of big data

Volume	The quantity of data
Velocity	Speed at which new data is collected
Variety	Data may be structured or heterogeneous
***Veracity**	Data can be noisy, incomplete, or wrong

Laney, 2001; Hurwitz et al., 2013

big data, *n*.:

Whatever doesn’t fit on your laptop

¯_(ツ)_/¯

More seriously...

- The definition of “big” depends on how the data is used and stored
- In practical terms, **“big data”** is differentiated by requiring coordinated processing by **multiple computers**
- Much of this class will focus on **distributed storage and computation**

Why this class?

- The tools are constantly evolving
- Odds are high that current software will be obsolete in a few years
- The underlying concepts don’t change so rapidly

⇒ Get proficient with concepts and current tools, and **learn to adapt**!

What should you get from this class?

- Familiarity with distributed storage and computation
- Appreciation for the technical challenges of big data
- **Understanding of when to use which methods and tools**

Your course staff for the semester

Instructor:	Brian McFee
Contact:	brian.mcfee@nyu.edu
Office hours:	Th 09:00-11:00 EST/EDT, http://bit.ly/dsga-1004-s22

Section leaders:

- Wed 17:55-18:45 - Xintong Li
- Wed 12:30-13:20 - Jack Zhu
- Wed 13:30-14:20 - Saumyaa Shah
- Wed 19:10-20:00 - Safwan Mahmood

+ Graders:

- Artie Shen
- Sanae Lotfi
- Bo Zhang

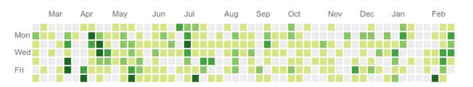
About the instructor...

(2018-) Assistant Professor of Music Technology and Data Science

machine learning algorithms and infrastructure for music and audio

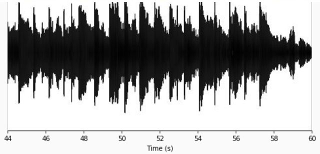
```
In [1]: import librosa

In [2]: 
```



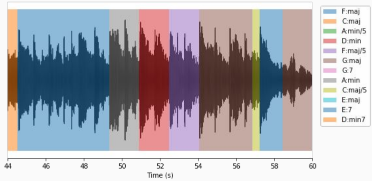
(2014-2018) Research fellow @NYU

Music and Audio Research Lab / CDS



(2012-2014) Postdoc @Columbia

Electrical Engineering / Center for Jazz Studies



(2012) Ph.D. in Computer Science, @UC San Diego

Similarity learning for music recommendation

Cross-modal learning

How does this class work?

- Read the syllabus!
  (No, seriously, it’s all in there.)

Lecture format

- Many of you may be unable to attend synchronously at some point
- We'll follow a **flipped classroom** format
- Lectures will be pre-recorded and posted in advance of our meeting time
- **Watch the videos on your own time before the class meeting**

Class meetings

- Class meetings will be used for discussions, Q&A, and group work
- We'll use Slido and work through problems together
- Staring at a screen all day is hard!
  - We might not use the full time, usually aiming ~1 hour

About jargon...

HadoopYARN
MesosKafkaCa
ssandraMongo
RedisSparkPar
quetNoSQLRD
BMSBeamHDF
SKubernetesHI

This subject matter involves a lot of obtuse terminology and buzzwords. **Don’t worry**.

I can’t keep most of the names straight either.

If terms are ever unclear, stop and ask for clarification.

Relatedly: some of you undoubtedly have more experience than others.

Be mindful of others and the environment we create in the classroom!

Readings

- Each week will have assigned reading, listed in the syllabus
  - Expect a book chapter, or 1-2 papers each week
- All materials will be available through **brightspace.nyu.edu**
- You're expected to do the reading **before class meets**
  - Learning works best when you first encounter new ideas on your own.
  - We can use the class time to clarify difficult or confusing concepts.
  - Give yourself time to do the reading -- **start early**!

Technology and resources

- All resources are available through **brightspace.nyu.edu**
    - Course schedule, assigned reading, etc...
- Lab assignments will be available via GitHub Classroom
    - If that’s a problem, we can make other arrangements

Grading

- 35% lab assignments
- 35% quizzes
- 30% final project

Lab assignments (35%)

- 5 ~bi-weekly programming assignments to be completed **individually**
- You'll get access to NYU's high-performance computing (HPC) cluster
- You have **2 slip days** to use however you like over the semester
  - After that, 20% penalty per day for late submissions.
  - No assignments will be accepted more than 5 days late.
  - Grading these assignments is not easy, please be mindful of the graders' time!

Quizzes (35%)

- 5 online quizzes, ~biweekly on Fridays
- Quizzes are open book, open note, but must be **completed independently**.
- You will have 1 hour to complete a quiz once you start.
- Lowest score is automatically dropped

Final project (30%)

- This will be an extended lab / programming assignment over 3-4 weeks, integrating several of the tools and methods that we'll cover.

- Due **5/13** (end of semester)
  - Slip days do not extend past 5/13!

- Details will be posted in April

Roadmap for the semester

1.	**01/24**	**Introduction**
2.	01/31	Relational databases
3.	02/07	Map-reduce
4.	02/14	Hadoop distributed file system
5.	***02/21***	***President’s day, no meeting***
6.	02/28	Spark
7.	03/07	Column-oriented storage
8.	***03/14***	***Spring break, no meeting***
9.	03/21	HPC and Dask
10.	03/28	Text and similarity search
11.	04/04	Reproducibility
12.	04/11	Recommender systems
13.	04/18	Graph algorithms
14.	04/25	Differential privacy
15.	05/02	Graphical processing units
16.	05/09	TBA

Let’s go!

I hope you enjoy the Spring semester!