CS 677: Big Data

Lecture 1

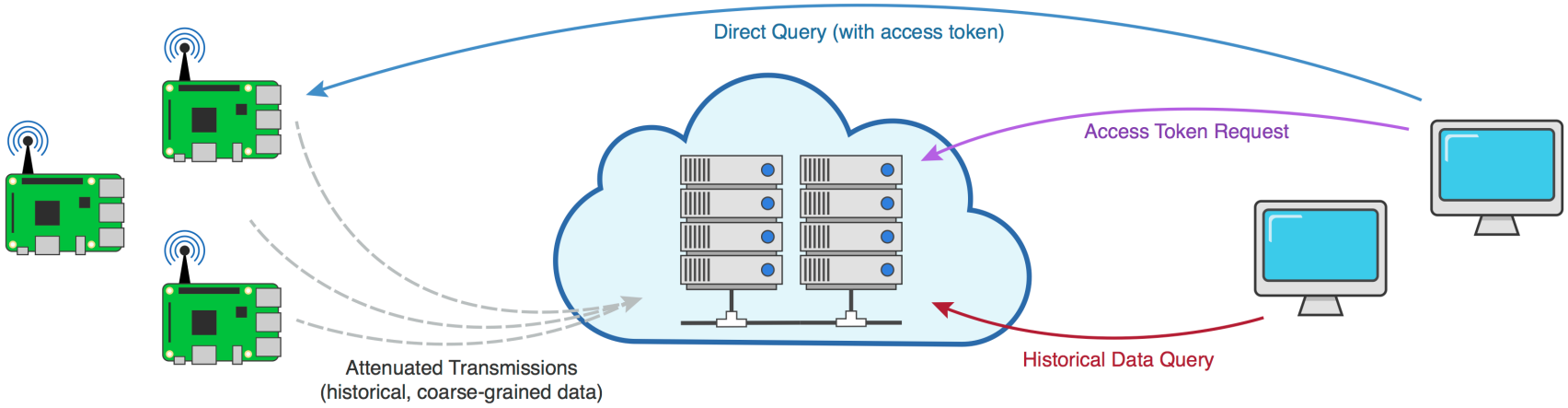
Welcome to CS 677!

- Looking forward to a great semester!
- *Lecture:*
  - *Lecture:* **Mon, Wed, & Fri** · 10:30 – 11:35 · **MH** 122
- *Website* (most course content will be here):
  https://www.cs.usfca.edu/~mmalensek/cs677

Staff

- **Instructor**: Matthew Malensek
  - mmalensek@usfca.edu
  - Office Hours: **M, W, F** 12:30pm – 1:30pm in **HR 407B**
- My background is in big data, distributed systems, and cloud computing
  - The goal of my research is to be able to process very large datasets quickly… by **not actually** processing them
    - We'll discuss a few ways to do that in this class

Edge/IoT/Fog Computing



Today's Agenda

- Syllabus
- What *is* Big Data?

Today's Agenda

- **Syllabus**
- What *is* Big Data?

Staying up to Date

- I will regularly update the schedule page on the course website
  - Upcoming topics, readings, due dates
  - Lecture slides, videos
  - URL: http://www.cs.usfca.edu/~mmalensek/cs677
    - Let's check out Week 1…
- If you need help, there will be Slack channel!
- **Grades** will be posted on Canvas
- Project submissions: GitHub

What We'll Discuss

- What's going on in the world of big data (research)
- How to build your own fault-tolerant distributed storage system
  - Modeled after production systems used by Google and Amazon
- How to build your own distributed computation engine
- How to use popular big data analysis tools such as Hadoop and Spark
  - We'll get some experience visualizing data and using machine learning models

Books

- There is no required book; we'll be reading research papers instead
- If you'd like background on related topics, these books can be helpful:
  - *Distributed Systems: Principles and Paradigms* by Andrew S. Tanenbaum and Maarten van Steen
  - *Big Data: Principles and best practices of scalable realtime data systems* by Nathan Marz
  - *Designing Data-Intensive Applications* by Martin Kleppmann

Course Structure

- We'll cover an assortment of algorithms, system designs, and overall "best practices" for working with extreme-scale data
- You will build big data systems and use big data tools on big datasets
  - So bigly big
- We'll read and scrutinize recent developments in big data research

Grade Distribution

- Projects: 40%
- Research Papers: 20%
  - Discussions: 5%
  - Presentation: 15%
- Quizzes: 25%
- Participation & Labs: 15%

Projects

- We'll have ~4 projects this semester
- Two are **system building** projects
  1. Build a distributed storage system
  2. Build a distributed computation engine
- Two are **analysis** projects
  1. Given a dataset, build distributed computations to analyze it
  2. Analyzing a dataset that **you** find interesting
- We'll primarily use Go and Python for these

Research Papers

- We will read several research papers throughout the semester. There are two parts to these assignments:
  1. Reading the paper and participating on the class slack channel
  2. Presenting a research paper to the class. This is a group assignment (~3 students per group) that involves reading the paper, performing a thorough analysis of its strengths and weaknesses, and then explaining the concepts to the class.

Quizzes

- Given roughly every four weeks
- Helps us go back and review what we've covered
- The strategy to do well is:
  1. Take note of concepts that we cover in class. These are most important.
     - Sometimes questions are even revealed in advance…
  2. Review the slides.
  3. If something isn't clear, review the lecture videos and book chapters

Grading

Score	Grade
100 – 93	A
92 – 90	A-
89 – 87	B+
86 – 83	B
82 – 80	B-
79 – 77	C+
76 – 73	C
72 – 70	C-
69 – 67	D+
66 – 63	D
62 – 60	D-
59 – 0	F

Policies

- Assignments are due at 11:59 pm on the due date
- Late lab and research paper assignments are not accepted.
  - In most cases, missing one of these has little to no impact on your grade
- Project deadlines are a bit more complicated…
  - …

Today's Agenda

- Syllabus
- **What *is* Big Data?**

What exactly does “big” mean?

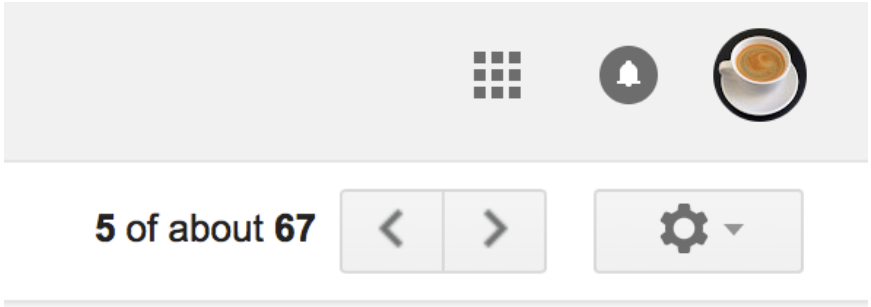
- Google was processing about ~25 petabytes of data per day back in 2010
  - That might as well be a century ago in the tech world
- Amazon has **exabytes** of customer data stored on S3
- Humanity is generating tens of zettabytes per year while connecting billions of new devices
- AI companies are rabidly **scraping** the web, discord, torrents, usenet, and your grandma’s collection of crossword puzzles to feed their models

Big Data [1/2]

- Humans and machines generate **massive** amounts of information every single day
- Ever bought anything on Amazon.com?
  - They recorded your searches, when you visited the site, what products you hovered your mouse over (and for how long), what you clicked on, bought, etc., etc…
  - Everybody is doing this. Not just the huge companies!
- Sensors and network connectivity are **cheap**
  - Radars, GPS, climate sensors, LIDAR, satellites, self-driving cars, smart phones and watches

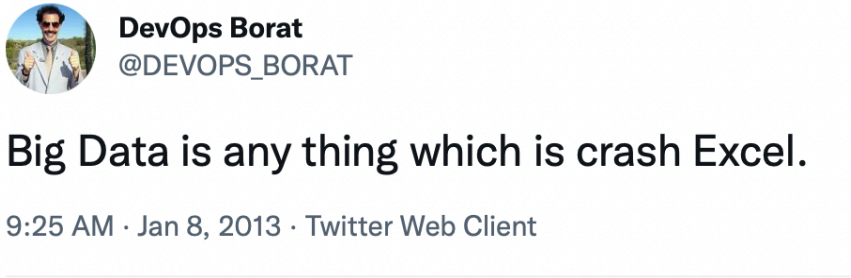
Big Data [2/2]

- Both humans and machines generate massive amounts of information
- The way you deal with “Big Data” is…
  - … you don’t!
- If I hand you a petabyte’s worth of anything and you have to inspect every item, it’s already game over



So… what IS it?!

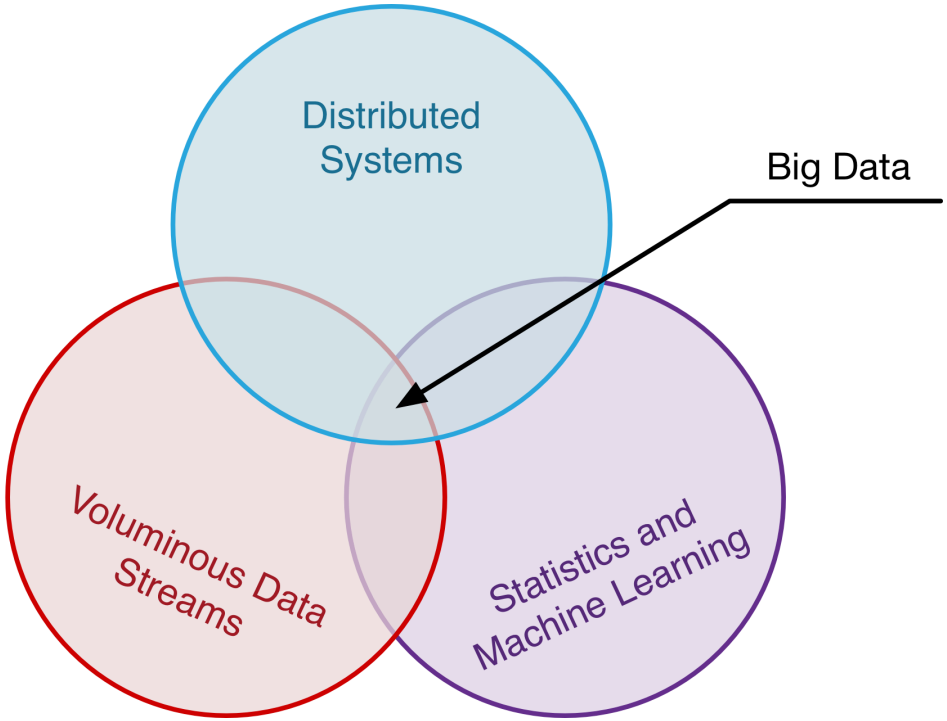
- Some folks just have a whole lot of data
  - Does that count?
- Big files? Lots of tiny files?
- Anything that you can use Spark to analyze?
  - Hey! We haven't talked about that yet…



So… what IS it?!

- In this class, we'll define anything that requires more than a single machine to analyze **efficiently** as a "big data problem"
- Another way to look at it is the problems cannot be solved efficiently using traditional algorithms
  - i.e., you can't just start looping over every data point. We have to come up with a different approach.

Our Focus


Distributed Systems
Big Data
Voluminous Data Streams
Statistics and Machine Learning

Wrapping Up

- Big Data is all about:
  1. Scale: accomplishing tasks that are just too large/intensive to do on one or a few machines
  2. Insight: extracting knowledge from the data (or in business terminology, “extracting value”)
- Coming up next:
  - A few big data case studies
  - Research papers
  - Go(lang)