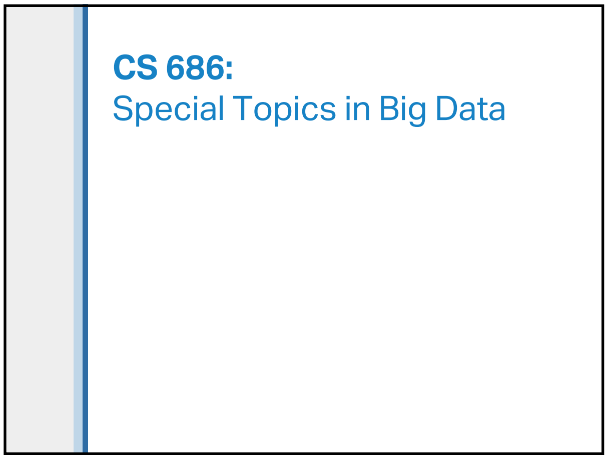
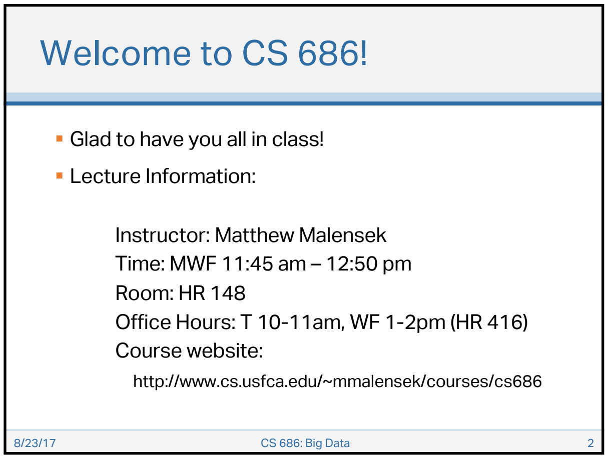


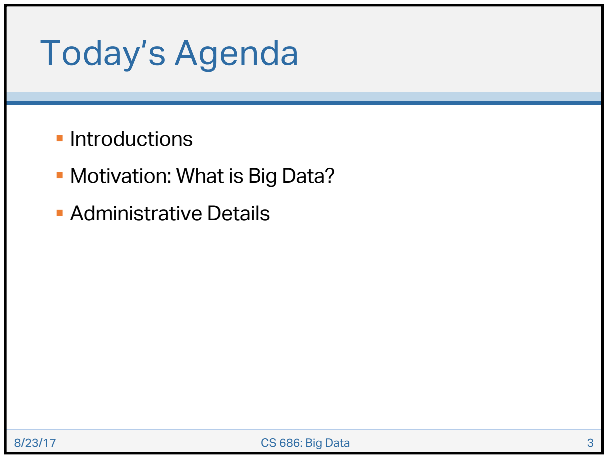
CS 686:
Special Topics in Big Data



Welcome to CS 686!

- Glad to have you all in class!
- Lecture Information:

Instructor: Matthew Malensek
Time: MWF 11:45 am – 12:50 pm
Room: HR 148
Office Hours: T 10-11am, WF 1-2pm (HR 416)
Course website:
http://www.cs.usfca.edu/~mmalensek/courses/cs686



Today’s Agenda

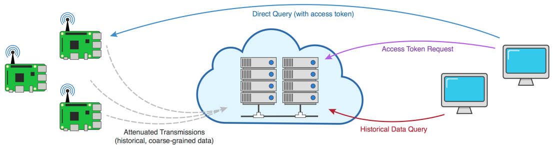
- Introductions
- Motivation: What is Big Data?
- Administrative Details

Today's Agenda

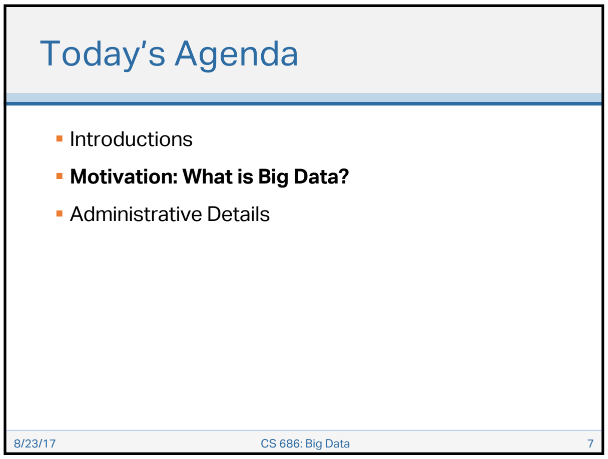
- **Introductions**
- Motivation: What is Big Data?
- Administrative Details

A Bit About Me

- My research is on big data, distributed systems, cloud computing, and data science


Direct Query (with access token)
Access Token Request
Historical Data Query
Attenuated Transmissions
(historical, coarse-grained data)

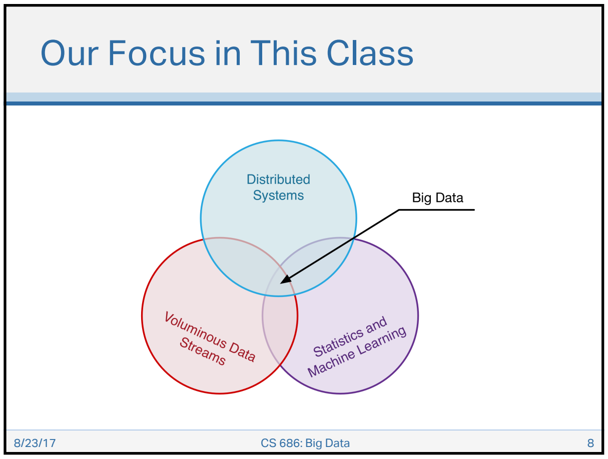
A Bit About You!



Today’s Agenda

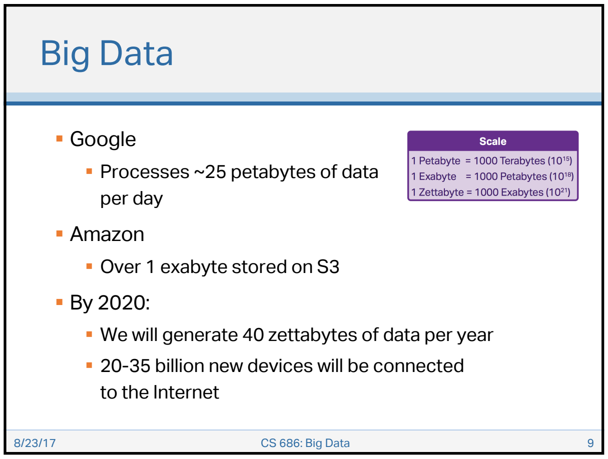
- Introductions
- **Motivation: What is Big Data?**
- Administrative Details

8/23/17 CS 686: Big Data 7



Our Focus in This Class

8/23/17 CS 686: Big Data 8

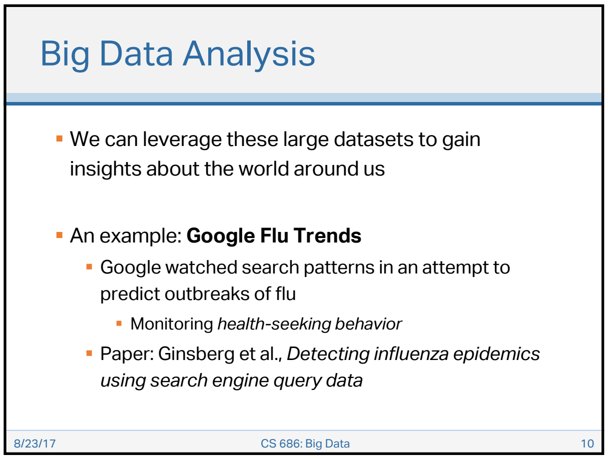


Big Data

- Google
  - Processes ~25 petabytes of data per day
- Amazon
  - Over 1 exabyte stored on S3
- By 2020:
  - We will generate 40 zettabytes of data per year
  - 20-35 billion new devices will be connected to the Internet

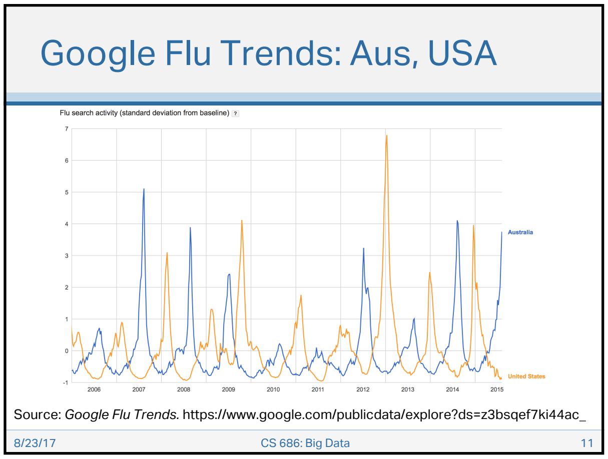
Scale
1 Petabyte = 1000 Terabytes (10^{15})
1 Exabyte = 1000 Petabytes (10^{18})
1 Zettabyte = 1000 Exabytes (10^{21})

8/23/17 CS 686: Big Data 9



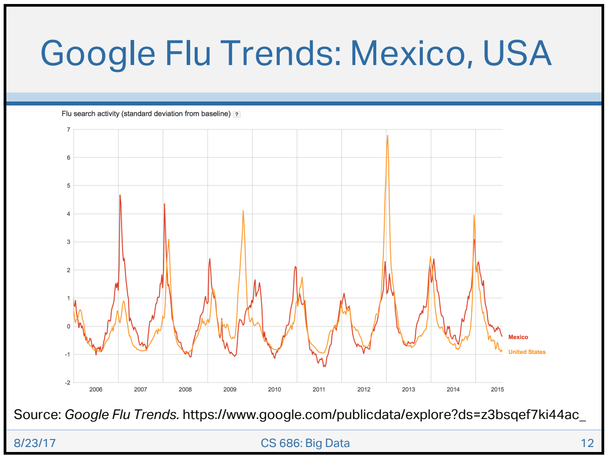
Big Data Analysis

- We can leverage these large datasets to gain insights about the world around us
- An example: **Google Flu Trends**
  - Google watched search patterns in an attempt to predict outbreaks of flu
    - Monitoring *health-seeking behavior*
  - Paper: Ginsberg et al., *Detecting influenza epidemics using search engine query data*



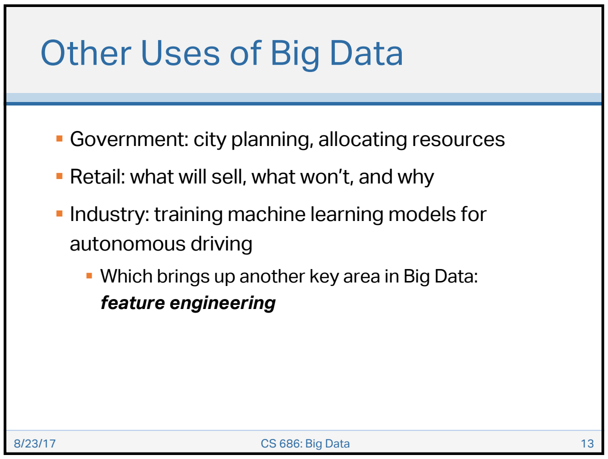
Google Flu Trends: Aus, USA

Source: *Google Flu Trends.* https://www.google.com/publicdata/explore?ds=z3bsqef7ki44ac_



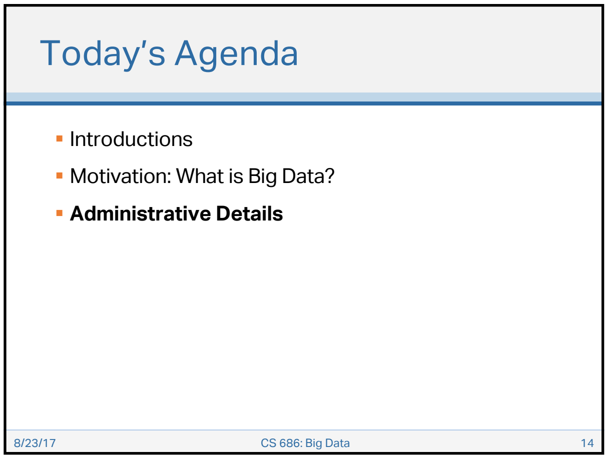
Google Flu Trends: Mexico, USA

Source: *Google Flu Trends.* https://www.google.com/publicdata/explore?ds=z3bsqef7ki44ac_



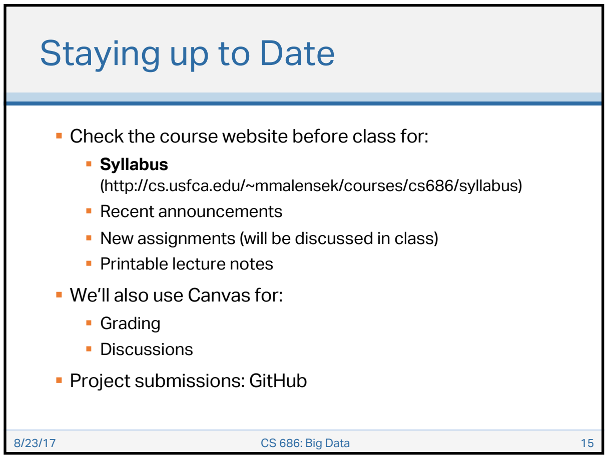
Other Uses of Big Data

- Government: city planning, allocating resources
- Retail: what will sell, what won't, and why
- Industry: training machine learning models for autonomous driving
  - Which brings up another key area in Big Data: ***feature engineering***



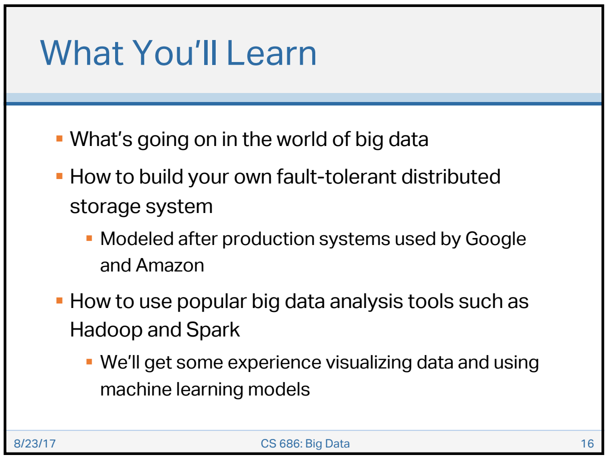
Today's Agenda

- Introductions
- Motivation: What is Big Data?
- **Administrative Details**



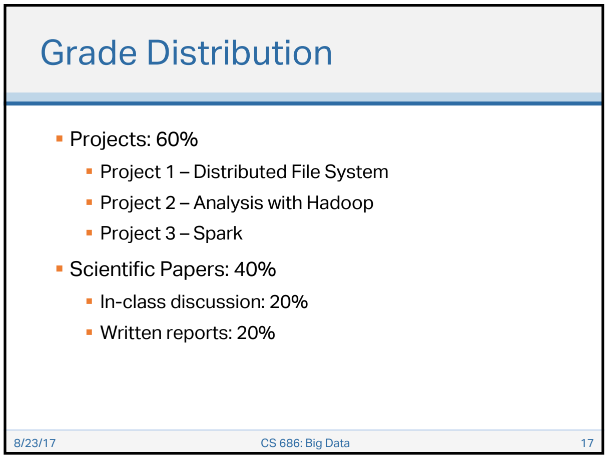
Staying up to Date

- Check the course website before class for:
  - **Syllabus** (http://cs.usfca.edu/~mmalensek/courses/cs686/syllabus)
  - Recent announcements
  - New assignments (will be discussed in class)
  - Printable lecture notes
- We'll also use Canvas for:
  - Grading
  - Discussions
- Project submissions: GitHub



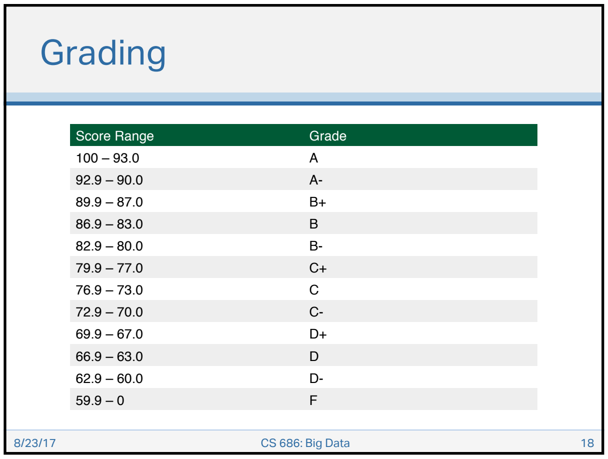
What You'll Learn

- What's going on in the world of big data
- How to build your own fault-tolerant distributed storage system
  - Modeled after production systems used by Google and Amazon
- 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



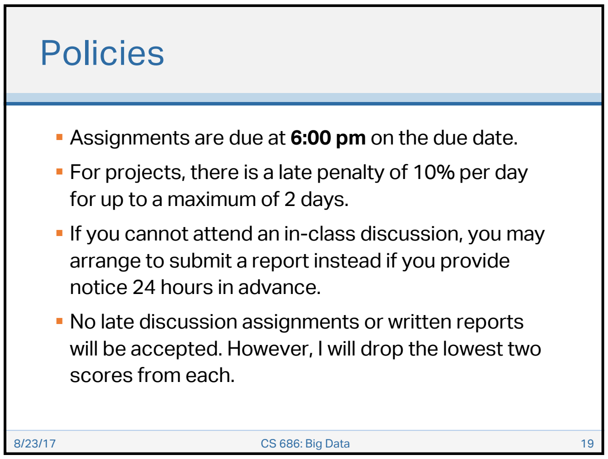
Grade Distribution

- Projects: 60%
  - Project 1 – Distributed File System
  - Project 2 – Analysis with Hadoop
  - Project 3 – Spark
- Scientific Papers: 40%
  - In-class discussion: 20%
  - Written reports: 20%



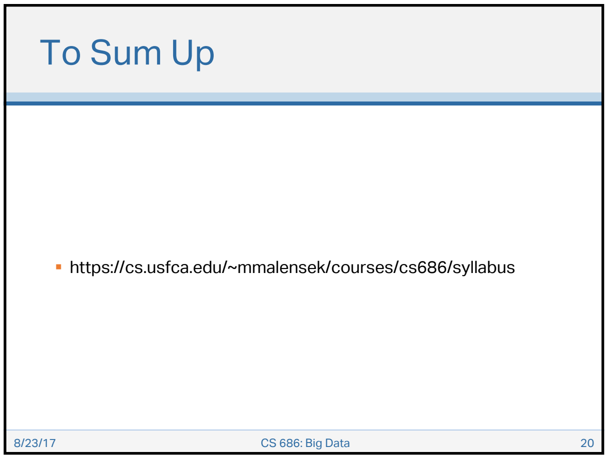
Grading

Score Range	Grade
100 – 93.0	A
92.9 – 90.0	A-
89.9 – 87.0	B+
86.9 – 83.0	B
82.9 – 80.0	B-
79.9 – 77.0	C+
76.9 – 73.0	C
72.9 – 70.0	C-
69.9 – 67.0	D+
66.9 – 63.0	D
62.9 – 60.0	D-
59.9 – 0	F



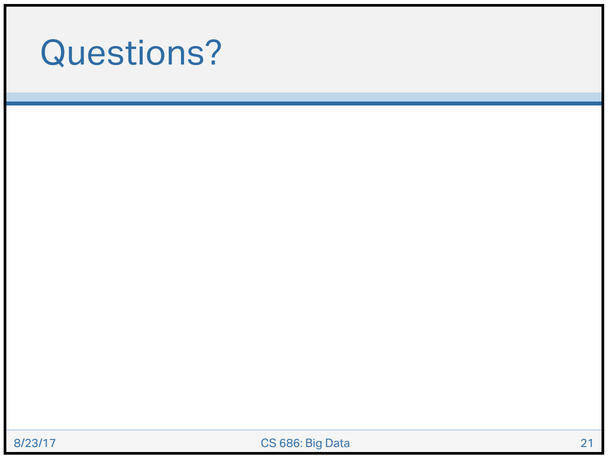
Policies

- Assignments are due at **6:00 pm** on the due date.
- For projects, there is a late penalty of 10% per day for up to a maximum of 2 days.
- If you cannot attend an in-class discussion, you may arrange to submit a report instead if you provide notice 24 hours in advance.
- No late discussion assignments or written reports will be accepted. However, I will drop the lowest two scores from each.



To Sum Up

- https://cs.usfca.edu/~mmalensek/courses/cs686/syllabus



Questions?