CS 677: Big Data

Lecture 1

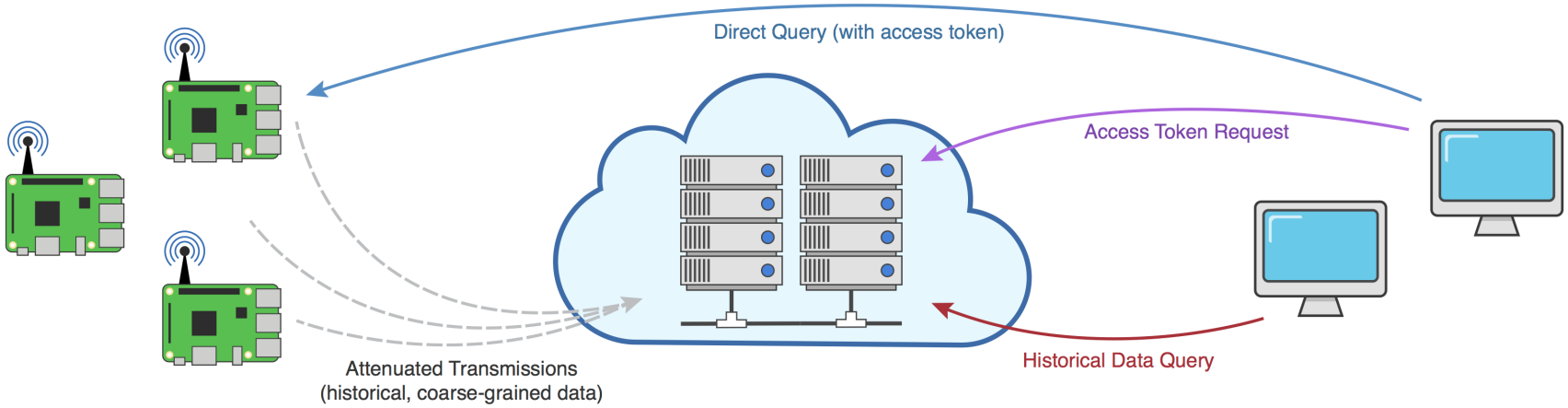
Welcome to CS 677!

- Looking forward to a great semester!
- *Lecture:*
  - **Tuesday & Thursday** · 9:55am – 11:40am · **LM 152**
- *Website* (most course content will be here): https://www.cs.usfca.edu/~mmalensek/cs677

Staff

- **Instructor**: Matthew Malensek
  - mmalensek@usfca.edu
  - Office Hours:
    - On Campus: **T, Th** 1:30pm – 2:30pm in **HR 407B**
    - Remote: **M, W** 9:30pm – 10:30pm **on Zoom**
- 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: CampusWire
- **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 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: 50%
- Research Papers: 20%
  - Discussions: 5%
  - Presentation: 15%
- Quizzes: 20%
- Participation & Labs: 10%

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 discussion board
  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.
- Project deadlines are a bit more complicated…

Today's Agenda

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

What exactly does “big” mean?

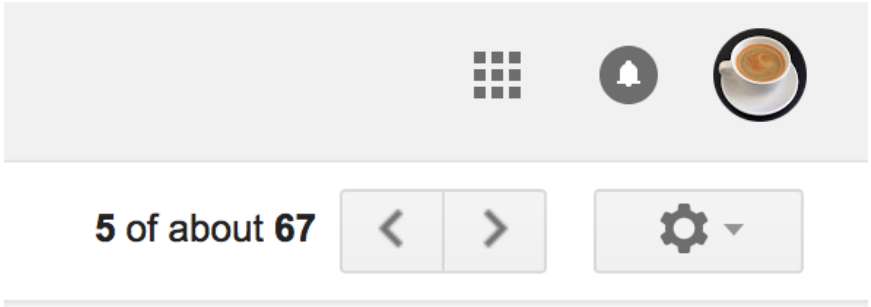
- Google was processing about ~25 petabytes of data per day back in 2010
  - That might as well be a century ago in the tech world
- Amazon has **exabytes** of customer data stored on S3
- Humanity is generating tens of zettabytes per year while connecting billions of new devices

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 Hadoop to analyze?
  - Hey! We haven't talked about that yet…

DevOps Borat
@DEVOPS_BORAT

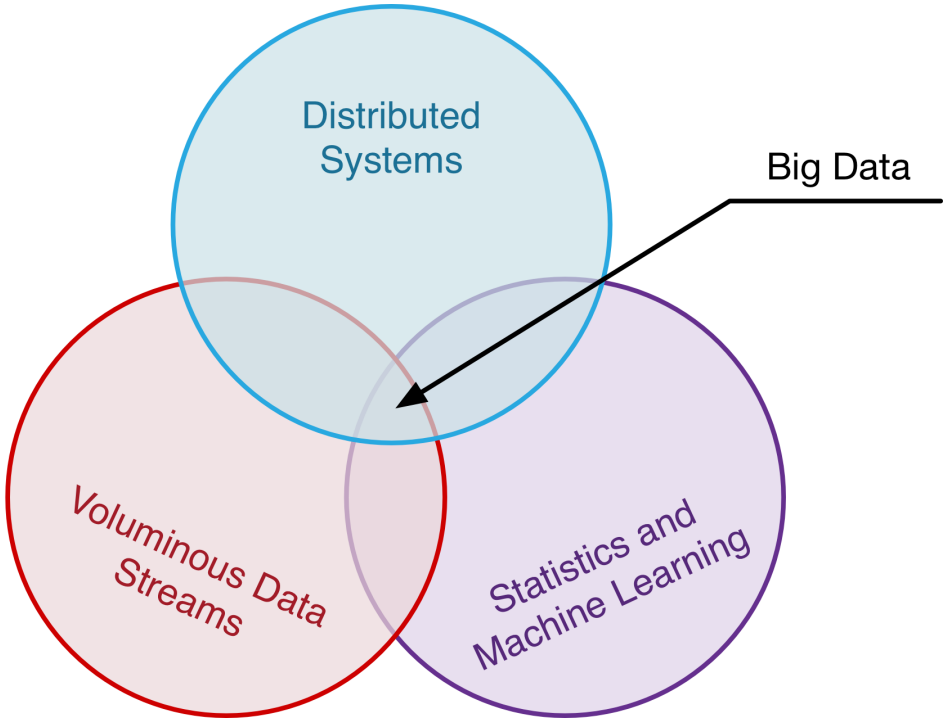
Big Data is any thing which is crash Excel.

9:25 AM · Jan 8, 2013 · Twitter Web Client

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

Big Data Analysis

- We can leverage these large datasets to gain insights about the world around us
- Let's look at a couple examples
  - Target's targeted advertising
  - Google Flu Trends

Big Data Case Study #1: Target

- The truth is, a huge amount of big data research is fueled by advertising
  - Companies want to be able to pinpoint your interests, recommend products, and get you to buy stuff
- Sometimes it doesn't work so well
  - I just bought a new TV, so Amazon assumes I probably want to buy a few more
  - Amazon also thinks I'm my wife, which is interesting…
- Let's look at a big data "success": Target

Target + Pregnancy

- Target wants to get expecting mothers hooked on their products before their bundles of joy arrive
- Things they track:
  - Customer ID number
  - Credit card
  - Name
  - Email address
  - A history of everything you have bought
  - Demographic information
- Note that Target buys some of these pieces of information: they don't have to collect it themselves

Detecting Pregnancy

- In the first 20 weeks, expecting mothers purchased lots of supplements (zinc, calcium, magnesium, etc.)
  - Morning sickness is common so there may be large quantities of ginger-based products purchased
- Most pregnant women buy large quantities of unscented lotion around the beginning of their 2nd trimester
- Nearing the end of the pregnancy: unscented soap, cotton balls, hand sanitizer, washcloths
- Target took this data, combined with other indicators, and produced a pregnancy score and estimated delivery date
  - And then they sent out coupons, of course!

Target + Pregnancy = Creepy

- A Minnesota father headed to his local Target to complain because his daughter received lots of baby-related coupons and offers
- His interpretation: Target is trying to encourage my high school-age daughter to get pregnant!
- The reality: she was already pregnant…

A "Happy" Ending?

- Target realized that their correct-but-creepy analytics was probably not the best idea
- Luckily, there was an easy fix: just mix enough random products in with the targeted ads
- Now the sheep ***cough*** customers won't be creeped out but will still "benefit" from our targeted ads!
- And in some ways, this story is almost too good to be true…

Pondering this a Bit More

- If **Target** can do this with fairly unsophisticated techniques… what can **Google** do with all the data they have on you?
  - Put on your tin foil hats
- Google knows what you search for, who you email, where you drive, what stores you shop at, who you send SMS to, and where you are at all times
  - Don't believe me? They can figure out your location just based on what WiFi networks are near your phone

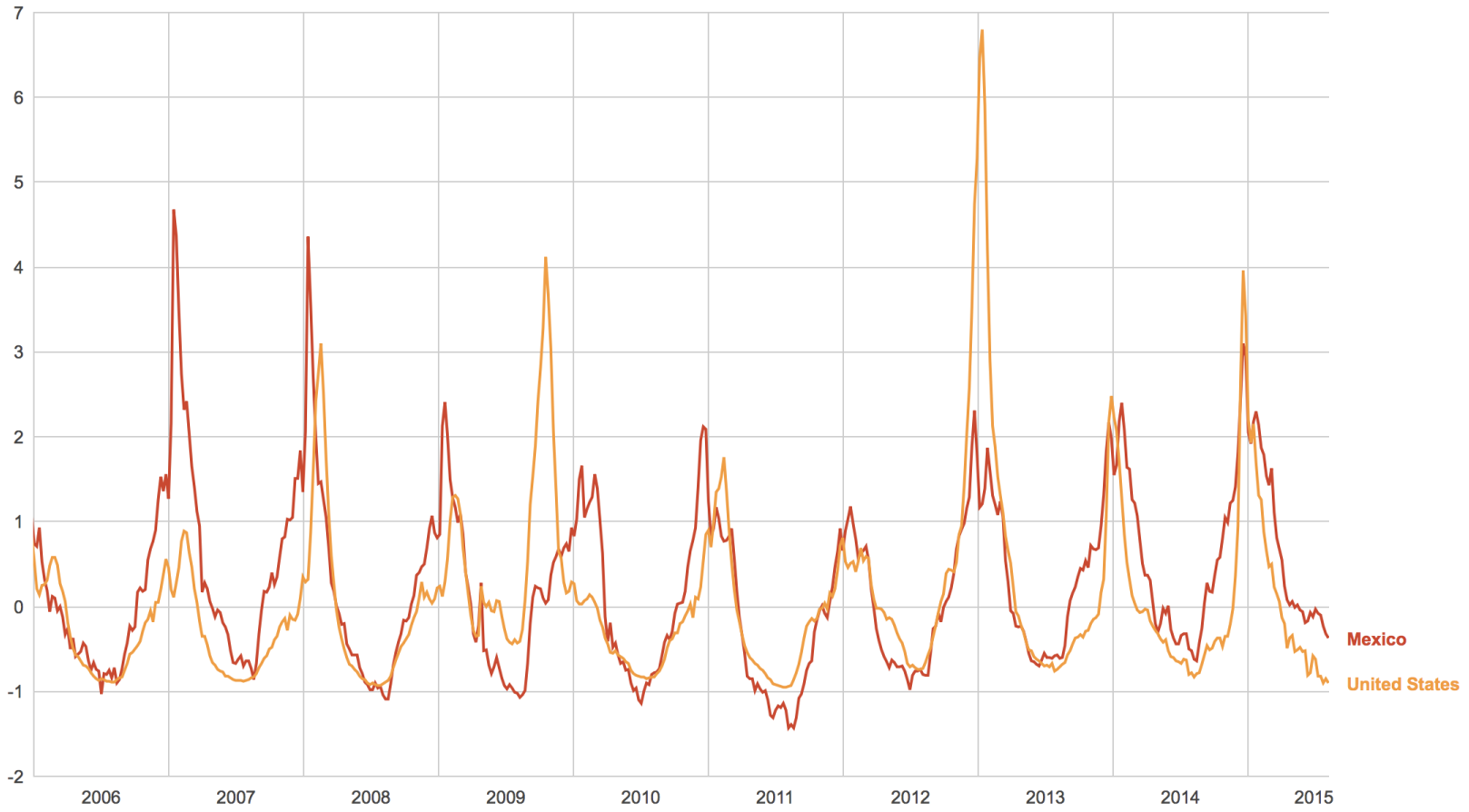
Big Data Case Study #2: Flu Trends

- An early example of big data analytics, Google used their search engine to predict influenza outbreaks
- 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*

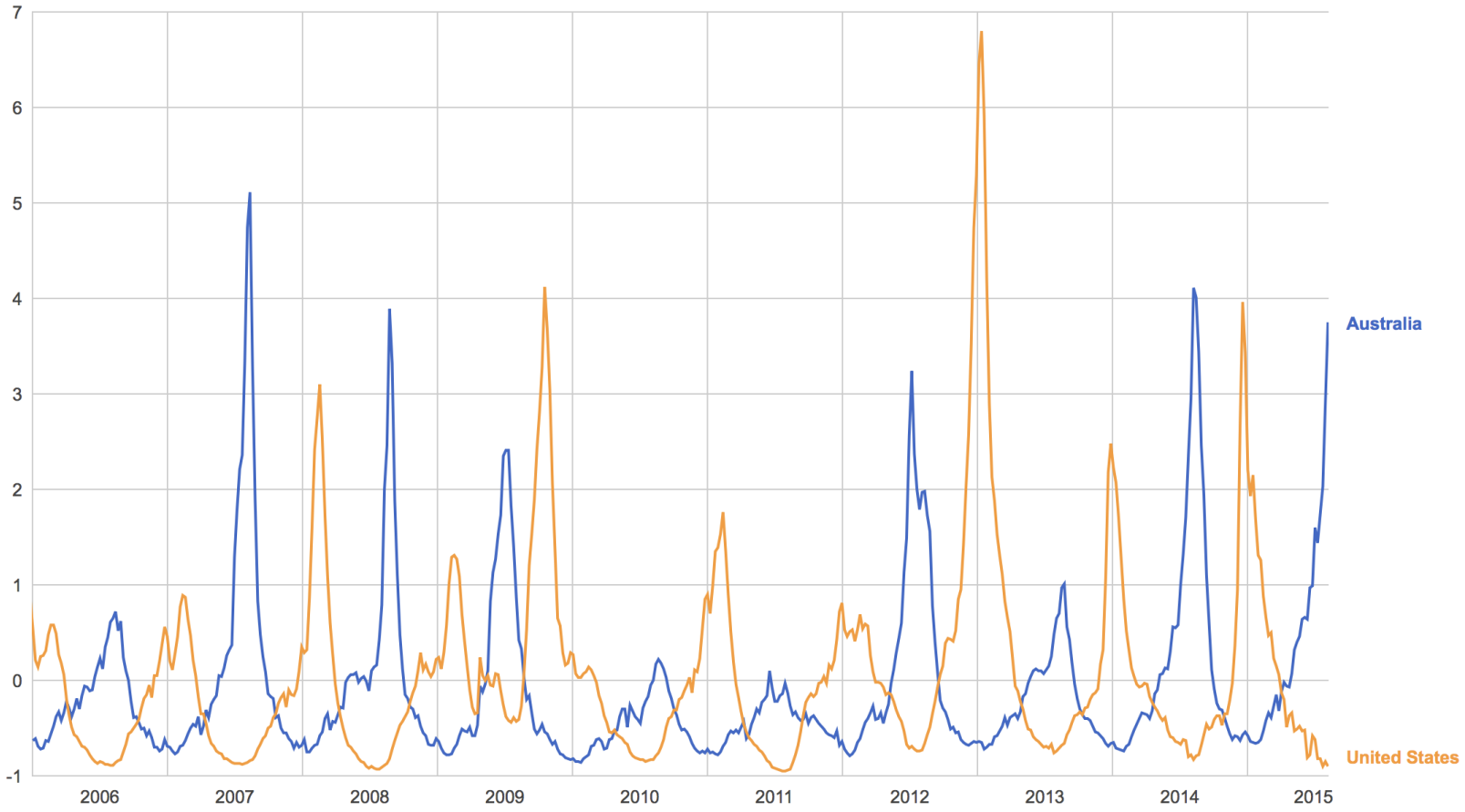
Health Seeking Behavior?

- Google is uniquely positioned to know fine-grained details about your daily life
- Or: they can judge you so much
  - Yeah, so what if I googled "Justin Bieber lyrics." I was only doing research!
- Uh oh, did you just Google "runny nose?"
  - Hmm, I see that 45 other people at USF just Googled flu-like symptoms… better sound the alarm
    - And 64 just googled "where do my tuition dollars actually go?"

Google Flu Trends: Mexico vs. USA



Google Flu Trends: Aus vs. USA



Outcomes

- GFT predictions were ~97% accurate, based on data from the CDC (Centers for Disease Control)
- People were pretty excited about this
  - Lots of positive media coverage
- So, yeah, Big Data = Awesome!
  - …but…

Flu Trend Failures [1/2]

- There was a huge amount of hype surrounding Google Flu Trends, but it has been discontinued
  - Why?
    - Well, Google does seem to discontinue everything…
- Initially, the model made good predictions
- Unfortunately, the model suffered from over-fitting
  - Even worse, changes in Google's search algorithm also changes the predictions

Flu Trend Failures [2/2]

- Ultimately Google created a fancy "winter detector"
  - Mind = Blown
- Simple time series models had better performance
- Worse, the engineers that created GFT weren't experts in epidemiology
  - Some search terms increased the likelihood of a flu prediction when they shouldn't have
  - CS folks are often guilty of falling into this trap
- So, this was a case of "Big Data Hubris"

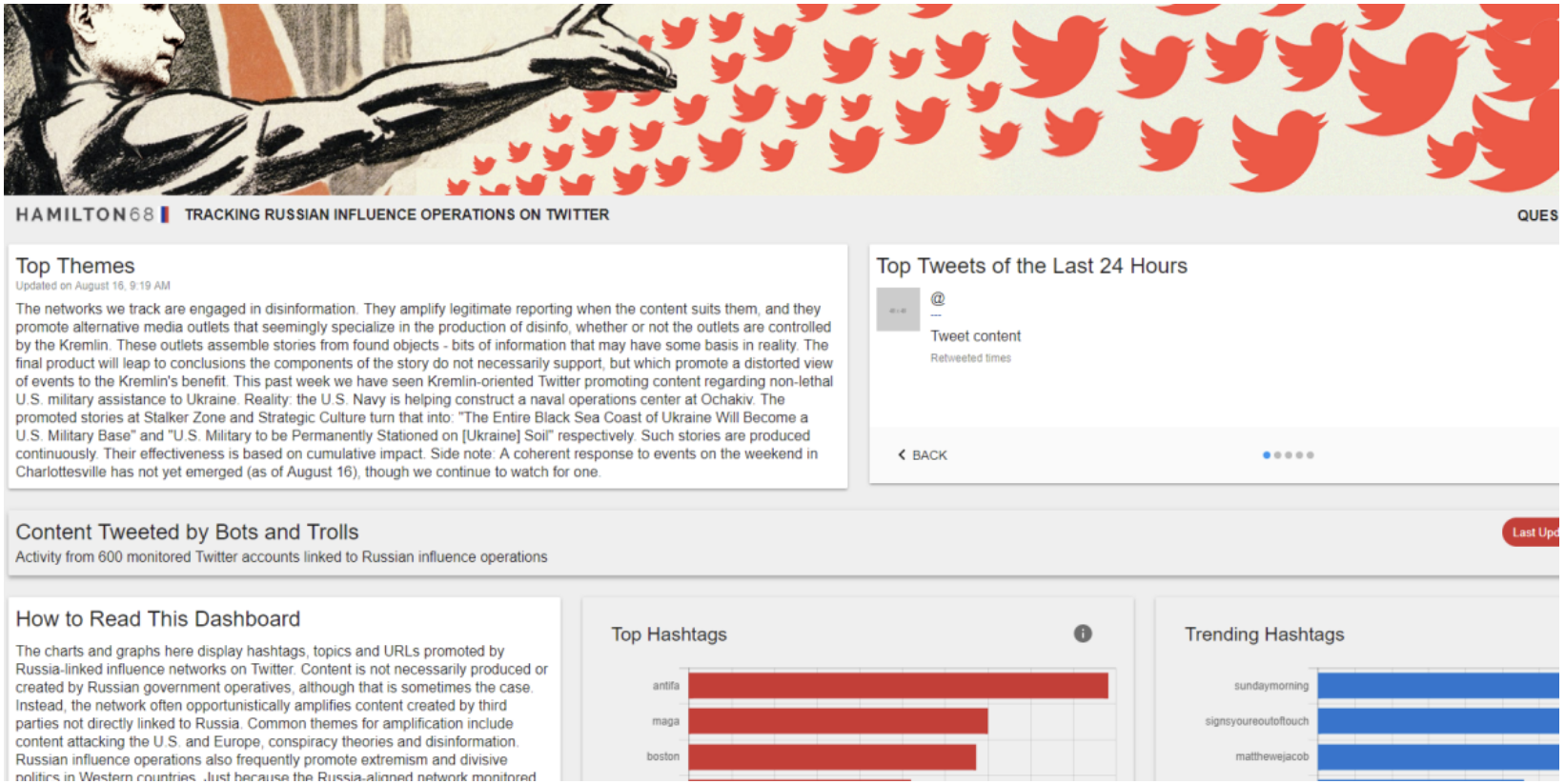
Big Data Case Study #3: Recommendations

- Netflix launched a competition to beat their movie recommendation algorithm
  - The prize: $1,000,000
- Netflix could predict within +/- 0.95 stars what you would rate a given movie
  - The goal: bring the error down to +/- 0.85 stars
- Took two years, combined hundreds of predictive models
  - Some problems are just really hard regardless of your datasets
- Never used by Netflix; after the transition to streaming, a bad recommendation is not as big of a problem

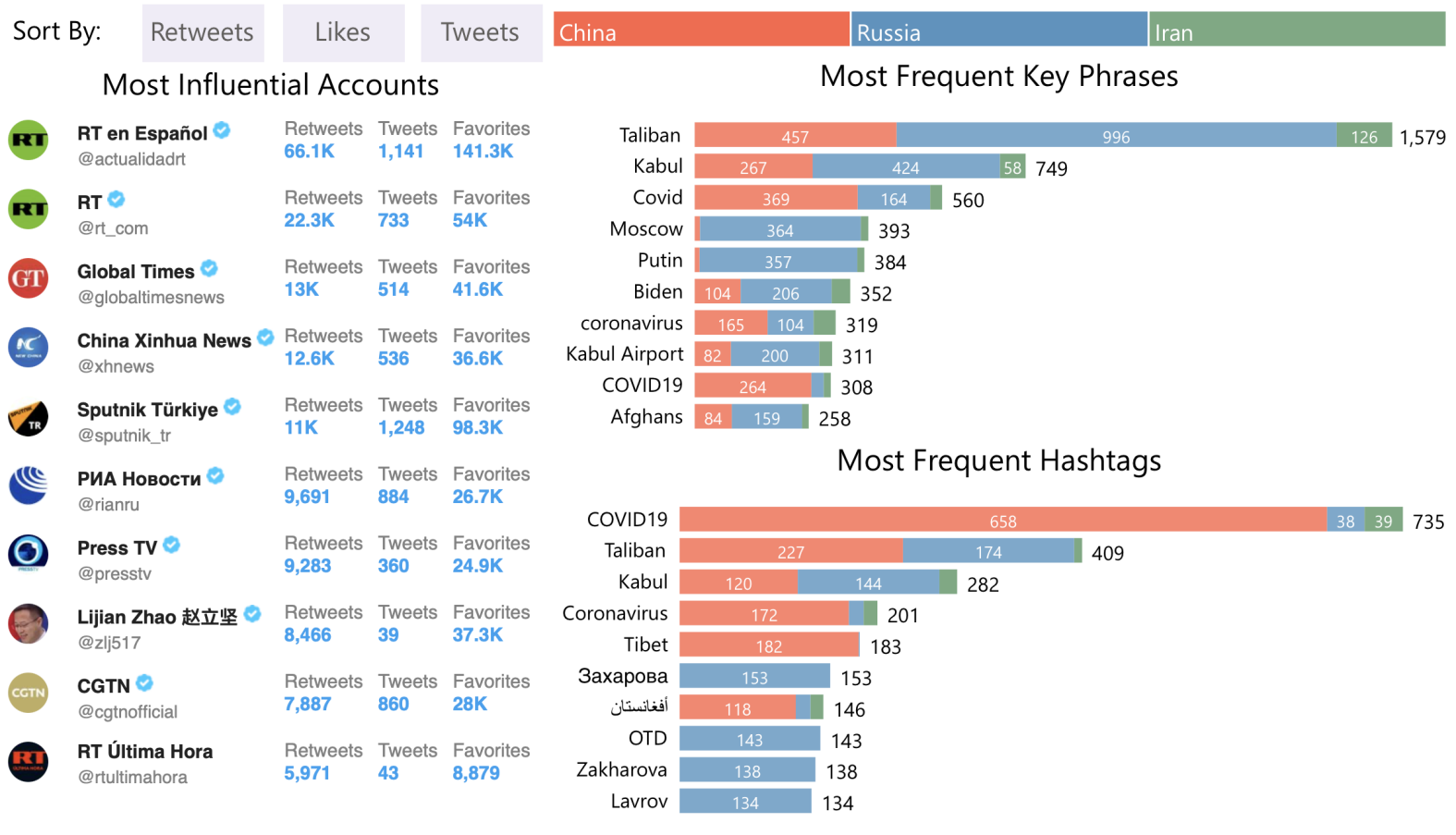
Big Data Case Study #4: Social Media

- Social media is a great source of data to exploit
- Spreading anti-vaccination info
- 5G Towers cause COVID-19
- We are ruled by the lizard people
  - No lizards were harmed in the making of this class
- Companies / governments can target your personal political beliefs
  - Companies want $$$, governments have $$$ but it looks bad when they spy on their citizens, so…

Big Data Case Study #4: Social Media



Big Data Case Study #4: Social Media



Facebook

- If you have a Facebook account, you can even view the "bubble" their algorithms have placed you in
  - See link on schedule page
- Reading things that agree with us makes us feel good
- These kind of social media issues are a bit similar to the early approaches toward security in CS
  - i.e., nobody really thought about it ;-)
  - Not just social media sites: news outlets run "submarine" pieces to influence you

Thinking Critically

- Big Data was a huge buzzword for a while, followed by some backlash – luckily we've hit a steady state in recent years
  - "Big data is like teenage sex: everyone talks about it, nobody really knows how to do it, everyone thinks everyone else is doing it, so everyone claims they are doing it…" – Dan Ariely
- Is Big Data the solution to everything? Nope!
  - Sometimes big data can actually help us create a "little data" solution
  - It's still a great way to explore and leverage patterns/insights

Other Uses of Big Data

- Government: city planning, allocating resources
  - Related: IoT, smart cities
- Atmospheric science, epidemiology
- Industry: training machine learning models for autonomous driving
  - And machine learning in general
  - Which brings up another key area in Big Data: **feature engineering**

Cleaning and Feature Engineering

- You may spend lots of time **cleaning** your data
  - Preparing it for analysis
- No data source is perfect. In fact, most are very, very imperfect
- Sometimes knowing what to **remove** or **combine** is just as important as collecting the data itself
- Feature engineering often requires domain knowledge

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")
- Hopefully these topics give you something to think about over the course of the semester
- One last order of business: Lab 0…
  - http://www.cs.usfca.edu/~mmalensek/cs677 (go to assignments page)