CS 677: Big Data

Big Data Case Studies

Lecture 2

Big Data Analysis

- We can leverage the large datasets we've been talking about to gain insights about the world around us
- Let's look at a few examples
  - Target's *targeted* advertising
  - Google Flu Trends
  - Movie Recommendations
  - Social Media

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 everyone in my house, 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*

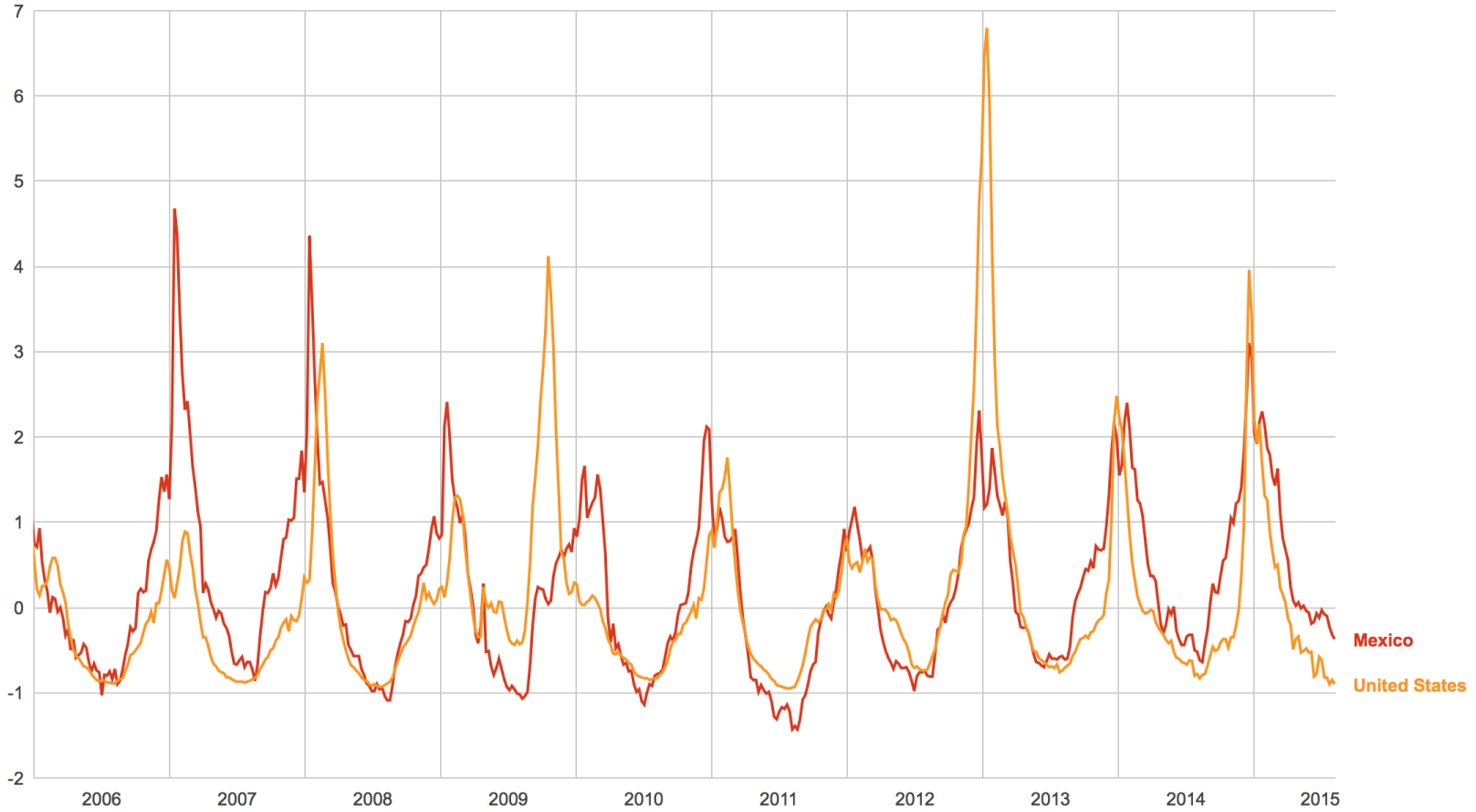
Health Seeking Behavior?

- 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?”

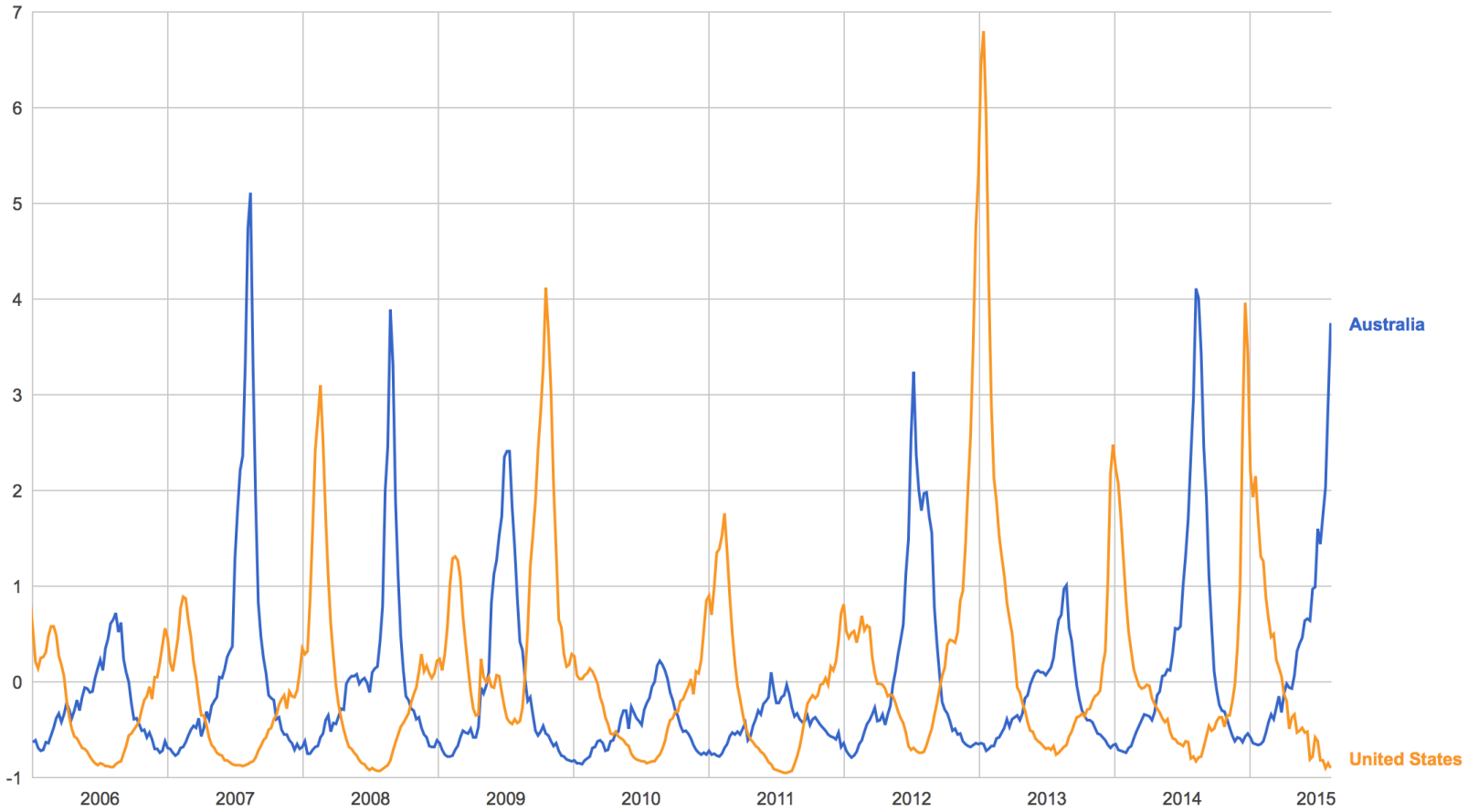
Reading break

- 📖 Let's read *Detecting influenza epidemics using search engine query data*
- With a teammate, split the paper in half.
  - One teammate reads the first half, the other reads the second
  - As you read, write down your **stream of consciousness**. It should be an append-only log. Don't delete/erase anything.
  - When you are done, try to figure out any missing pieces by talking to your teammate

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”

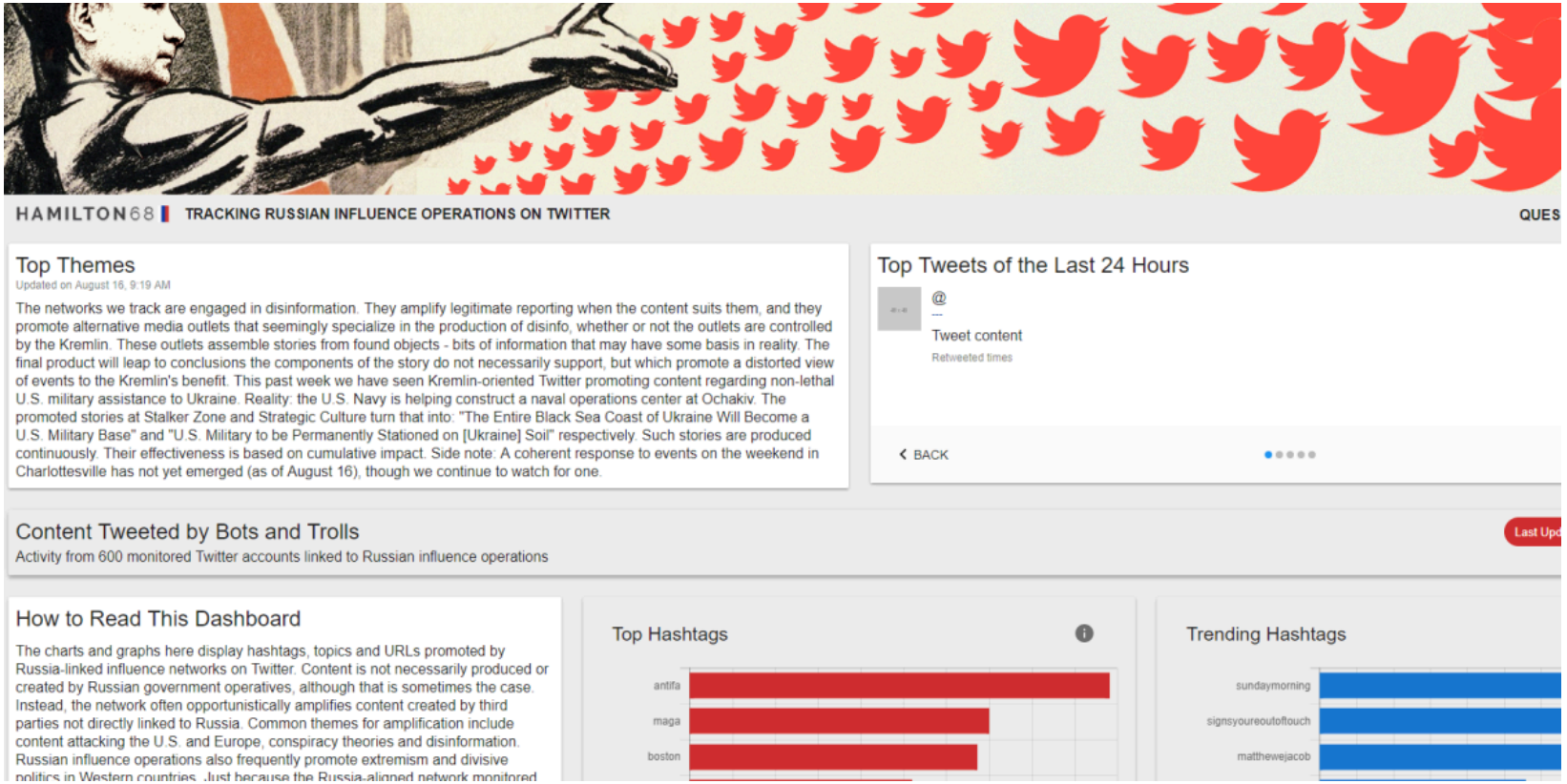
Big Data Case Study #3: Recommendations

- 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

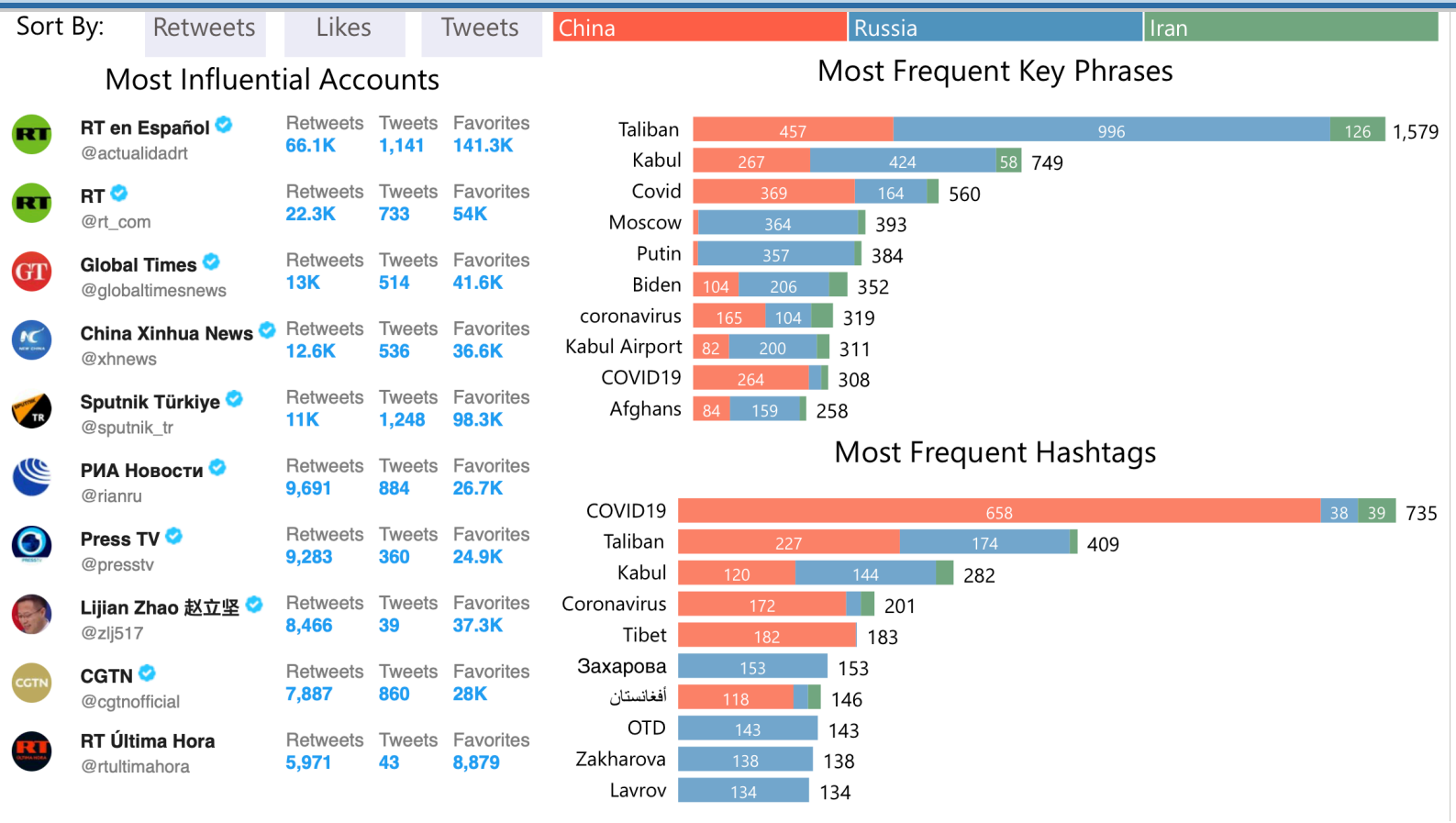
Big Data Case Study #4: Social Media

- 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

Thinking About Data

- To finish up, let's think about datasets
- What would you like to analyze this semester? Start out by determining:
  - What kind of data you need
  - What it might be able to tell you
  - Write this all down
- Then once you have a few options, it's time to do some research on actually collecting the information