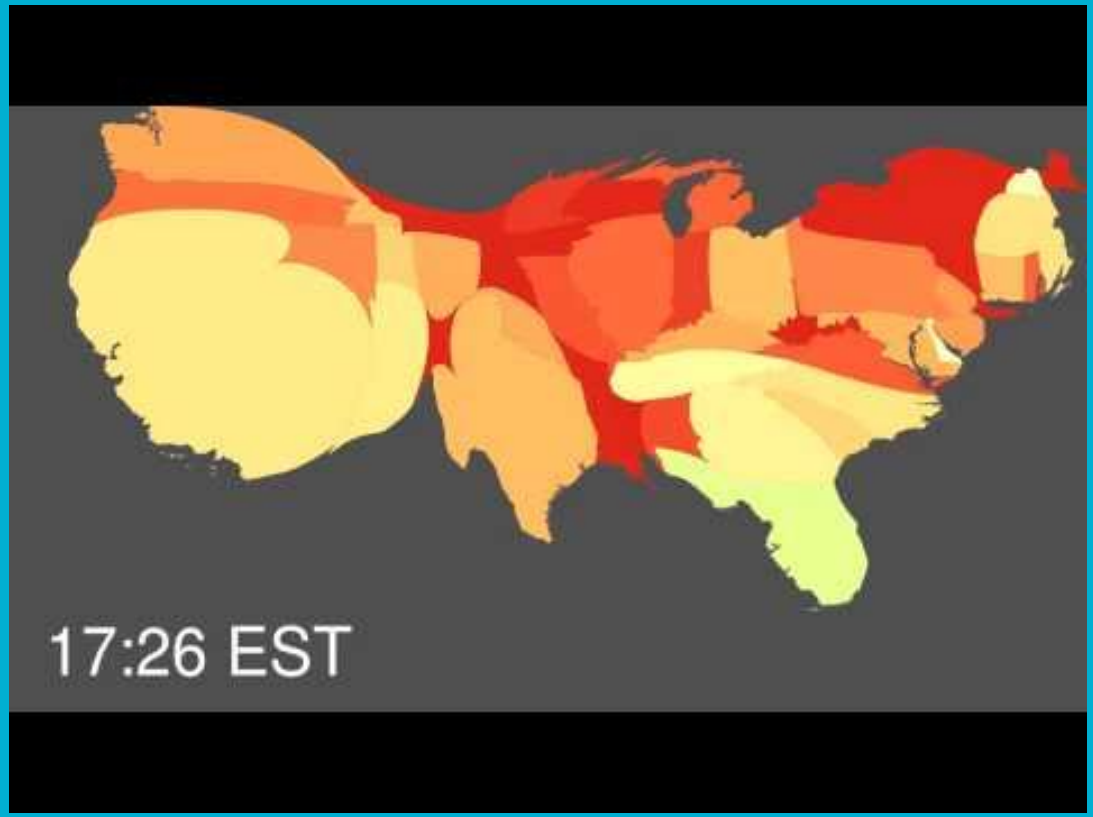
Social Media and Data Science

Rainstorm
Spring 2021
Instructor: Derek Lim

Question: What would you do if you wanted to know what emotions Americans felt at different times of day?

Take some time to brainstorm some ideas...

Pulse of the Nation: Inferring Mood from Twitter!!



Definitions: I bet you all know what social media is!!

Type in chat: Do you use any? Which one do you like most?

For today, **social media** is any online platform where people interact with other people and content generated by other people.

Examples: Facebook, Instagram, Twitter, TikTok, Pinterest

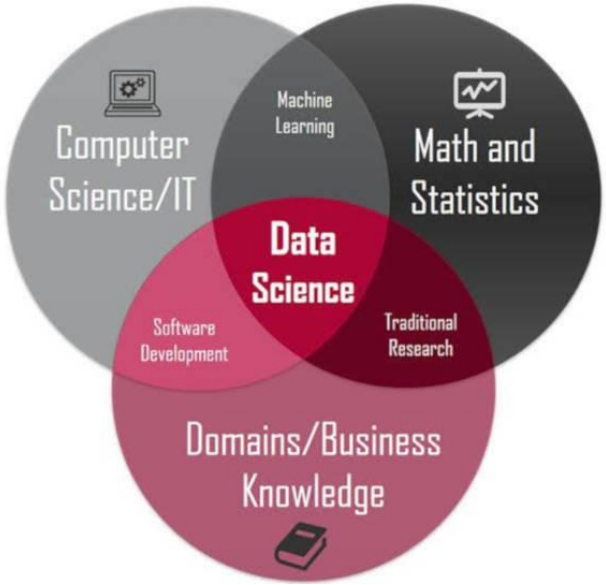
TikTok

Definitions: What is data science?

- Hard to define, involves using computers to learn and infer things from data

- One of the most in-demand fields in industry / research

- Useful for many applications: medicine, finance, economics, natural sciences, etc.



There are several big ideas in studying social media with data science

- **Idea A:** Big Datasets
- **Idea B:** Powerful computers and algorithms
- **Idea C:** Data through technology

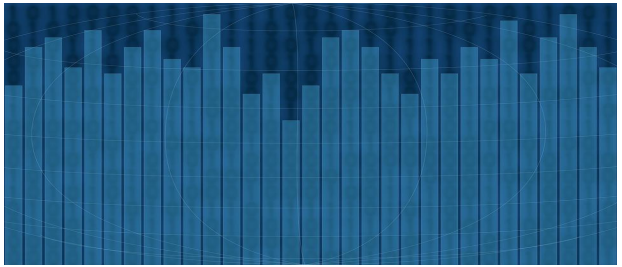
Idea A: Leveraging Big Datasets

- Question: can you guess how many people use Twitter each month?
- **Type your guess in the chat!**

- Answer: 192 million (as of Feb 2021) [1]

[1] https://www.businessinsider.com/twitter-earnings-q4-revenue-eps-new-users-2021-2

Idea A: Leveraging Big Datasets



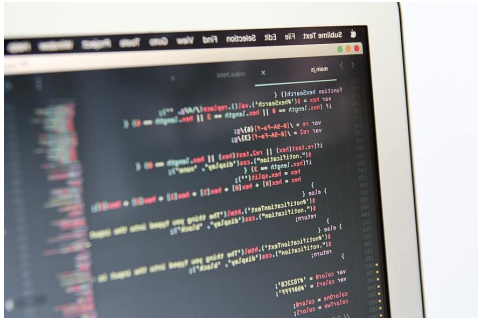
- **A lot** of data is generated by user activity on social media
  - E.g. profile information, interests, hobbies, behavioral properties

- Can get much more data than with telephone surveys, in-person studies
  - Lots of social science studies can have < 100 participants

- Can get data from users all across the world at all times
  - In-person studies mostly get data from locals during daytime hours

Idea B: Powerful computers and algorithms

- Every year, computers get faster (Moore's law)
- Specialized hardware allows even faster AI computations (GPUs)
- Smart, efficient algorithms developed by researchers and engineers allow learning from all types of data



Idea C: Data through technology

- Technology allows us to collect and access data that would be difficult or impossible to obtain otherwise

- Can connect to people from all over the world, at all different times

- When people do things online, they generate data (will discuss privacy later)
  - No data generated for most in-person things

Any Questions?

We will continue with some case studies of social media + data science ...

Question: Is it really a small world?

Case Study 1

- How connected are humans to each other?

- How many friends of friends do people have?
  - How about friends of friends of friends?
  - Friends of friends of friends of friends?
  - Friends^{10} ?

- Are most people friends of friends of … of friends of each other?

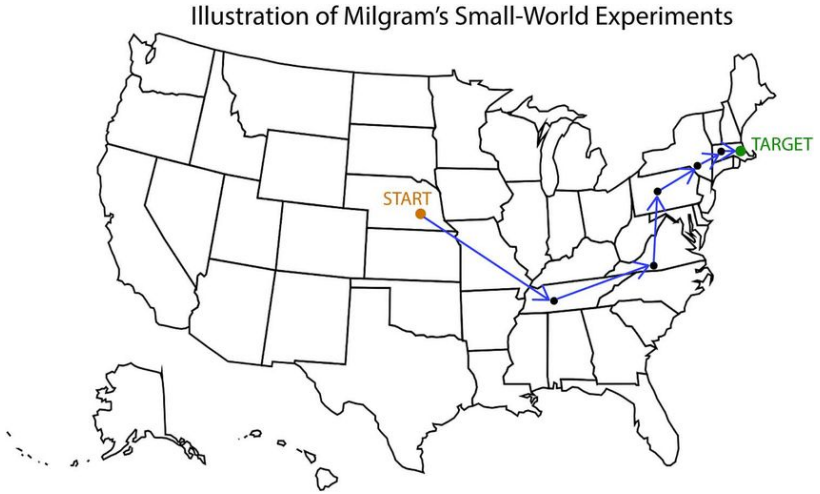
A Study from the 1960s: Milgram's 6 Degrees of Freedom

1. Fix one target person in Massachusetts
2. Ask 296 people in Nebraska and Boston to send a letter to this target person
3. Not given the target person's address, have to send the letter to acquaintances that may know the target person

- 64/296 letters reached the target!
- On average, letters traveled through about 6 acquaintances

Problems with the 1960s Study

- Only studies people in two parts of the USA, no people of other countries
- Selects 296 people, only 64 letters reach the target person
  - Only 1 target person
- People may not choose optimal path



Two 2012 studies with Facebook data!!

- Uses new computational tools (**Idea B**) and new large-scale data (**Idea A**) from a social media site (**Idea C**)
- Takes Facebook social network, and computes paths along friendships between different people
- Much larger experiment: (720 million users at the time)
- Studies connectedness of users all over the world

The world is really well-connected on Facebook!

- **Type in chat:** For a person with 100 Facebook friends, how many friends of friends do you think they will have?

- **Answer:** In 2012, about 27,500 unique friends of friends!!
- The study found that on average, there are **four degrees of separation** between two people on Facebook
  - Most pairs of people are friends of friends of friends of friends of friends!!

Question: How do people change the way they speak in different communities?

Case Study 2

- How do people talk differently around family, different friend groups, in sport teams, playing video games?
- How do people change their language when first entering a new friend group / community?
- How do people change their language as part of a community over time?

Past studies are uncertain on how language use changes

- Sociolinguistics: adult language stability assumption says that speech patterns of people do not change much in adulthood
- Evidence against adult language stability: people may change their language as they age
- Evidence against adult language stability: all people of community may adopt some language changes
    - E.g. many older people know fairly recent internet slang

A 2013 Stanford study uses online beverage review sites to study this!

- Data from two large beverage review communities with millions of posts (**Idea A, Idea C**)
  - Many users have written more than 100 reviews!

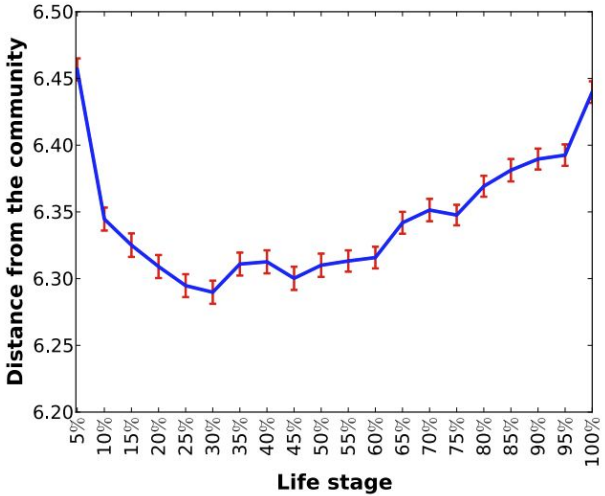
- Develop a formula for measuring how close the language of a post is to the language used by the community (**Idea B**)

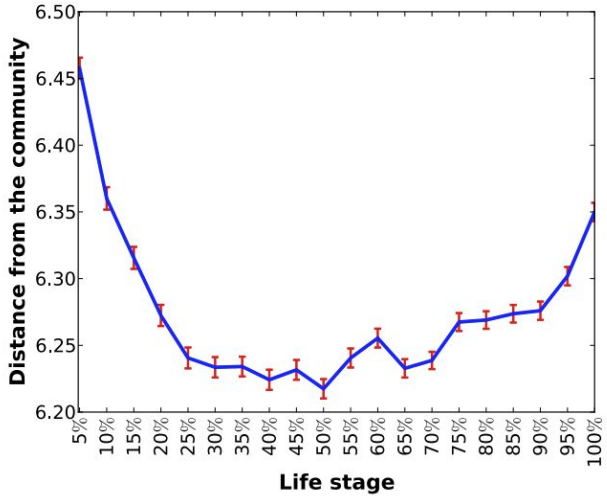
- Use this formula to study how user language use compares to the language of the community

Users are first unlike the community, then adapt their language to the community, then stop changing much

x-axis: portion of users' time on site that has passed

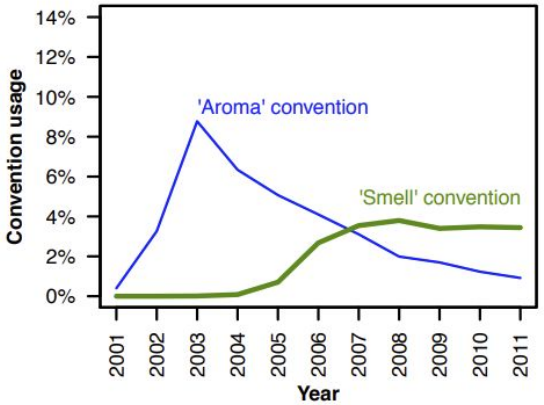
y-axis: how dissimilar users' speech is to community (lower value means more similar)



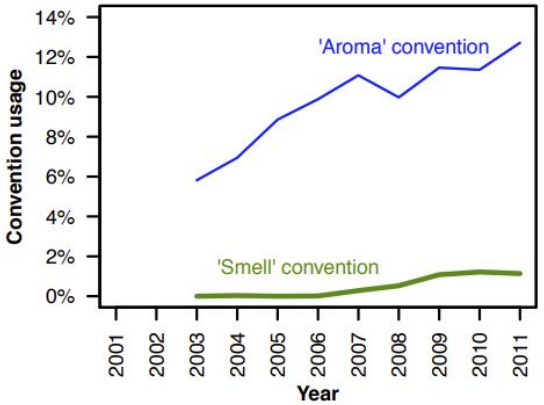


Change in language of smell discussions

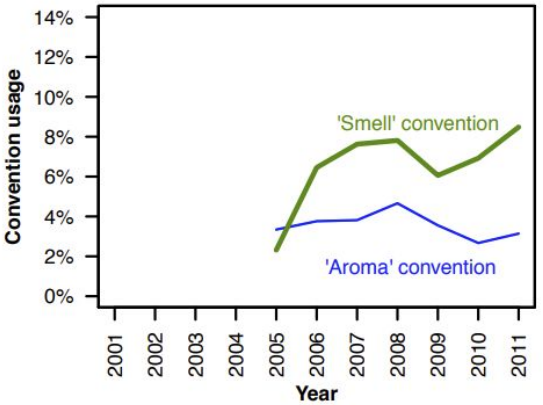
- Before 2005, users used “aroma” for discussion of smell, but then using “smell” became more popular
- Old users still use “aroma”



(a) ‘Aroma’ was the dominant convention by 2003, but it was supplanted by ‘S’ (for ‘Smell’) around 2007.



(b) Users who joined in 2003 hung on to the ‘Aroma’ convention of their youth.



(c) Users who joined in 2005 were more receptive to the emerging ‘S’ norm.

This evidence supports the adult language stability theory in online communities

- When a user is an “adult” on the site (have spent a lot of time), they may change their language less

- Who would have thought to use data from beverage reviews!

- Their formulas and analysis can be used for other online communities.

Any Questions?

We will conclude with some final considerations

Quick tour: what else has been studied in social media + data science?

- So much work on every popular platform!

- (My own work): how do users on genius.com (previously called rapgenius) decide what lyrics to annotate / when to annotate them?
- How can we identify fake news / misinformation?
- How does the layout of a website change what songs get popular?
- How do posts become viral?

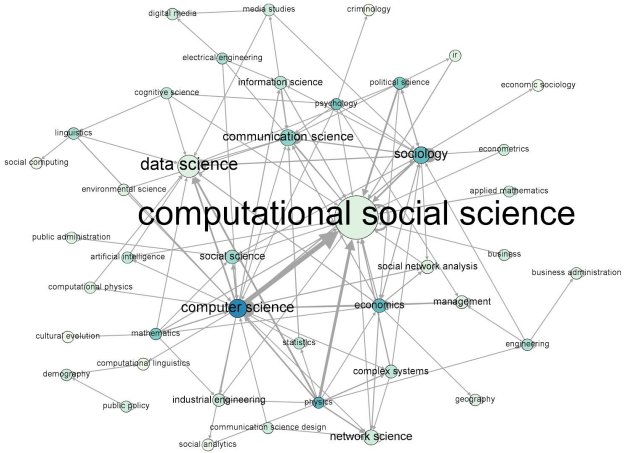
Ethical considerations are deeply important when dealing with personal data and making decisions

- User privacy is very important, should not leak personal information in datasets or analyses
- The conclusions that we draw can influence the world, we want the influence to be positive
- Different groups of people can be impacted differently
    - E.g. due to disability, national origin, race

What to study next (and keywords to Google search)?

- Main technical fields: computer science, statistics, math
- Technical subfields: machine learning, AI
- More specific fields: network science, sociology, other social sciences (e.g. economics, psychology)

- More buzzwords: computational social science, data science, information science



Conclusion + Thanks!

- Data science can be quite effective in analyzing data from social media
- **Ideas A, B, C** allow this combination to thrive
- Many examples show the power of combining social media and data science