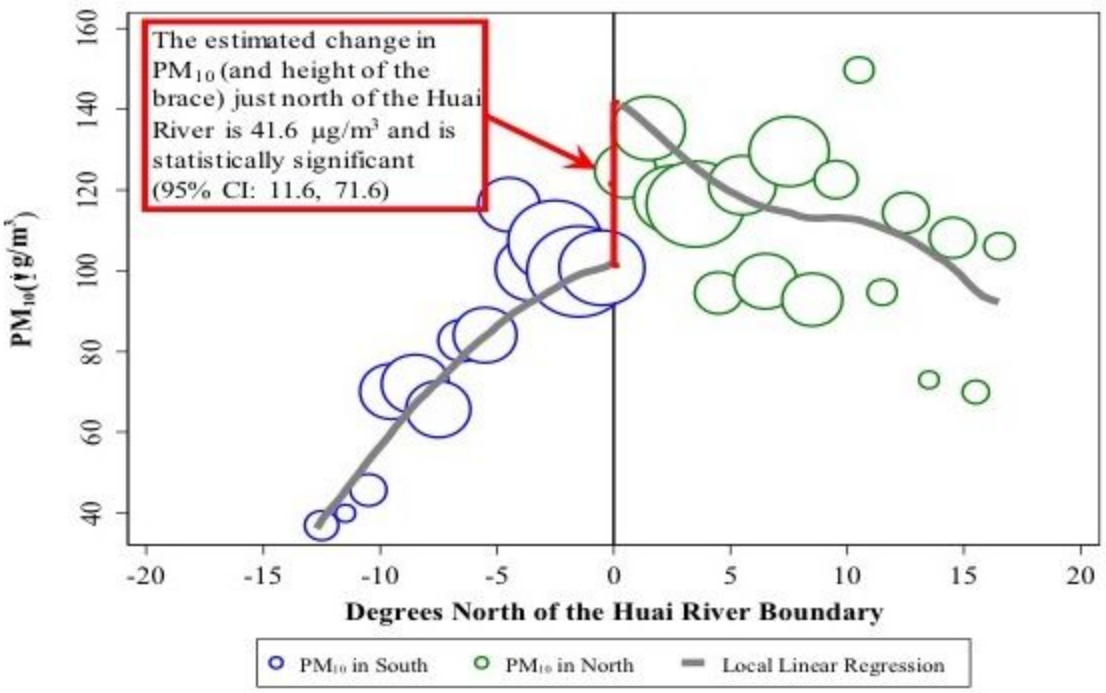
Slides here

Data is beautiful (and plentiful)

- Most academic journals require/encourage people to post data for public access
- Example of beauty:
    - Use Netvizz, Gephi to visualize social networks of Somali people living in Eastleigh.
    - More links among Somalis than among other people; social networking is like exchanging business cards
    - Demonstrates **Eigenvector centrality**: recursive network importance ranking (you are important because your friends are important)

Data is insightful

- Example: Does pollution in China matter?



-
    - Greenstone made a heatmap of PM 10 particulates in China
    - We see correlation between charts of life expectancy and pollution
        - North of the Huai heating (by coal) was subsidized.
        - Particulate matter levels increase as you rise to the North; statistically significant
        - Similar trends for life expectancy (similar shape, more life expectancy in the south)
        - Caveats: heat might save lives, might be other facts of living in the South

Data is powerful, part 1

Does anyone care about data?

Example: Changing regulation in India

- Background: Collaboration with the Gujarat Pollution Control Board in India
    - In India, an inspector and a third-party auditor inspects regulated polluting firms
    - Auditor has an interest to keep the relationship going
    - Because auditing was so corrupt, firms sued the government (essentially a tax on the firms, since the data was bad and unused)
- Proposal: change incentives:
    - Pay into a pool of auditors which is administered by the government.
    - To prevent cheating, introduce control of random backchecking (check again after a few days)

Data is powerful, part 2

- Experiment: introduce backchecking for all firms
    - For firms that are eligible for audit, randomly divide into treatment and control firm
        - We divide randomly to prevent adverse selection :P
    - Treatment: auditor is paid by a pool and backchecking is used to monitor
    - Control: used the usual system
- Result: treatment distribution similar to backchecks, but not to control
    - Conflict of interest leads auditor to cheat on data reported to the government
    - Experiment showed solution and changed the policy
    - And, the treatment group actually reduced their pollution!
    - Timeline: 2 years to prove, and more to get the policy changed...
    - Fun fact: you can collect data from firms (which don't have rights like humans) as long as it's protected

Data can be deceitful, part 1

Example: What affects autism?

- **Fabricated data:** Lancet publication that agent in measles vaccine causes autism
- **Correlation:** strong correlation between cases of autism and glyphosate (used to protect GMO plants); both series are strongly trending upwards. But perhaps not causated:
    - Perhaps more diagnosis is correlated with increased plant chemicals/agriculture
    - Many upwards trends: organic food sales also correlated with autism growth

Data can be deceitful, part 2

Less trivial example: Strong correlation between secondary school enrollment and GDP

- People use this data to increase funding for schools
- One interpretation: economic growth relies on human capital
    - In general, 1 year of education increases earnings by 7 to 8% in most countries
    - And, this benefit compounds (more education will share, etc.): externality of education
- But, perhaps this is not causal (another graph of non-logarithmic shows no relation). Why?

    - Reverse causality: perhaps GDP explains enrollment (wealthier people don't need children to work)
    - Third factor: perhaps government supports education and also malaria

Correlation v. causation

Correlation is not causation, and a causal narrative doesn't necessitate causality either.

Be careful of causality. If you search long enough, you'll find correlations: like,

- Revenue generated by arcades correlates w/ computer science doctorates in the US
- # of people drowned by falling into a pool & # of films Nick Cage appeared in

How do we extract something truly meaningful?

What you will learn in this class

- Probability: how do we model processes?
- Statistics: how do we summarize and describe data?
- Econometrics, machine learning: how do we uncover patterns?