Announcements:

- **Validity Reflection presentation due next week (6 min video, see template)**

Data Science by Example

CSE481DS Data Science Capstone

Tim Althoff

PAUL G. ALLEN SCHOOL
OF COMPUTER SCIENCE & ENGINEERING

Recap: Data Science Objectives

1. Formulate a research question
2. Identify a dataset with which to answer the question
3. Design an analysis process (next)
4. Consider construct, internal and external validity
   - Remember that more data doesn't necessarily help

Recap: Data Science Process

- **Plan** your own project along these stages
- When learning about other projects **pay attention potential pitfalls** across all phases
- When working on your own project, **explicitly address each step and failure modes**

Brief Recap of Study Context & Main Results

How Physically Active Are We?

Physical activity is extremely important for health [Lee et al., 2012]. But we do not know how much physical activity people get!

According to WHO:

- 5-54% of Germans don't get enough activity
- No data for Switzerland and Israel

Health research limitations today:

- High cost, short-term, limited scale
- Biases from self-reporting

Wearable and Mobile Devices

69% adults own smartphones in developed countries
46% in developing economies (rapidly growing)

Wearable and mobile devices generate massive digital traces of real-world behavior and health

Activity Tracking

Tracking actions

- Steps (automatic)
- Runs
- Walks
- Workouts
- Biking
- Weight
- Heart rate
- Food
- Drinks
- And many, many others

azumio

Dataset Statistics

- Data from 2011
- 717,527 anonymized users
- Users from 111 countries
- 68 million days of steps tracking
  - **100 billion** data points (2TB), Minute-by-minute
- Focus on 46 countries with >=1,000 users
  - 32 high-income, 14 middle-income countries

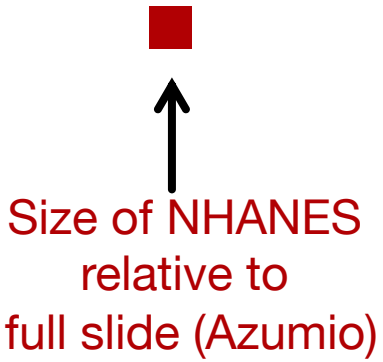
Today: 6M users, 160M days of activity, 800M actions tracked

Data in Context

- Our data: 68 million days of activity from over 700,000 individuals in 111 countries

1400x larger than largest existing gold-standard datasets:

- NHANES [Troiano et al., MSSE 2008]
- IPEN [Van Dyck et al., Int. J. Obes. 2015]



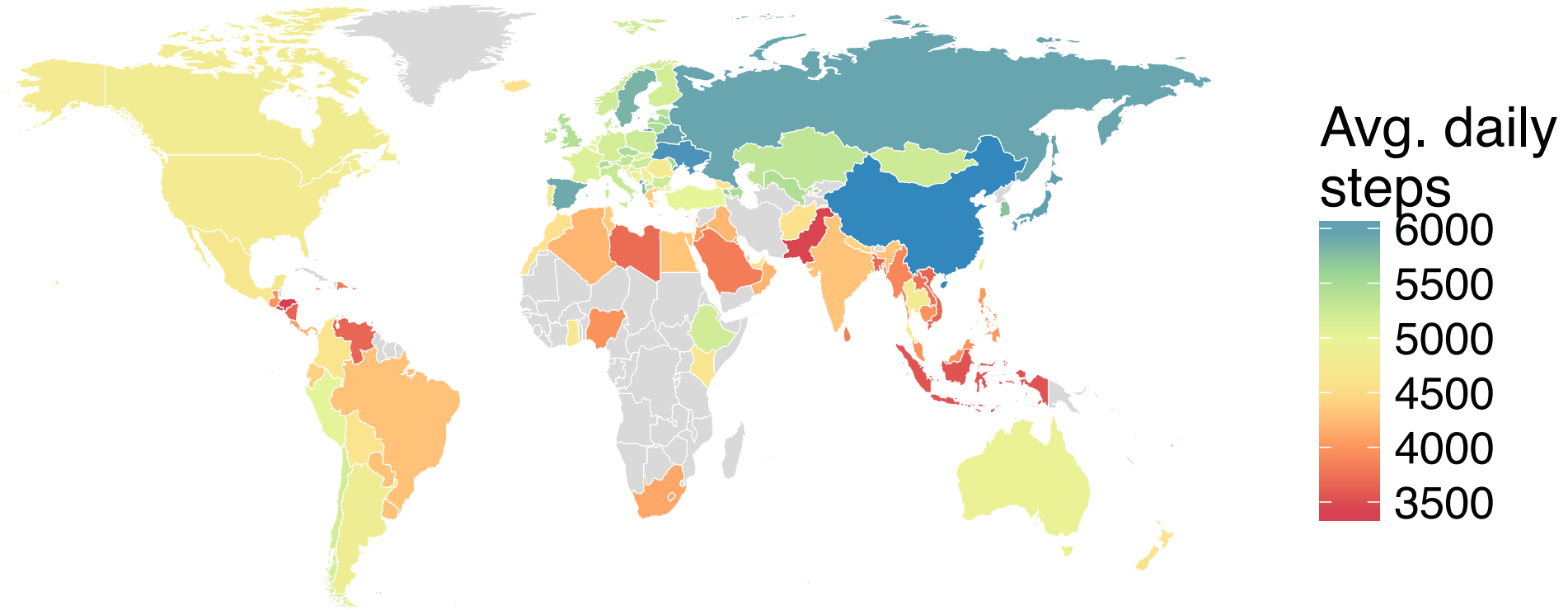
Population data available at:
http://activityinequality.stanford.edu/

Worldwide Activity

nature International weekly journal of science
Home | News & Comment | Research | Careers & Jobs | Current Issue | Archive | Audio & Video
Archive > Volume 547 > Issue 7663 > Letters > Article

Large-scale physical activity data reveal worldwide activity inequality

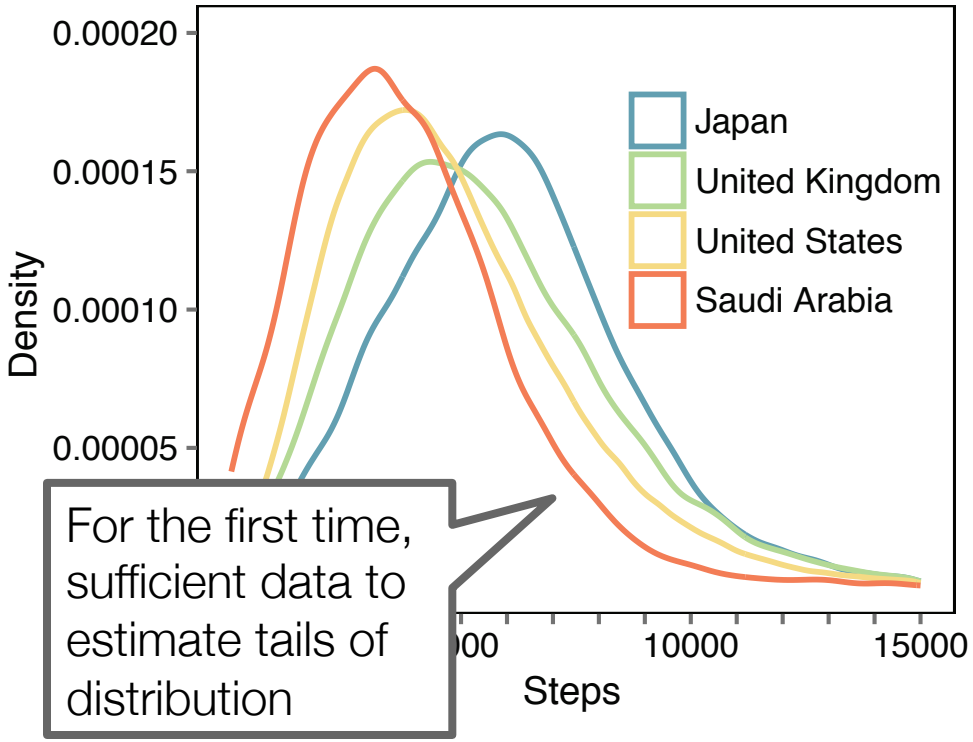
Tim Althoff, Rok Sosič, Jennifer L. Hicks, Abby C. King, Scott L. Delp & Jure Leskovec



But, how is activity distributed within the population?

Result 1: Inequality of Physical Activity

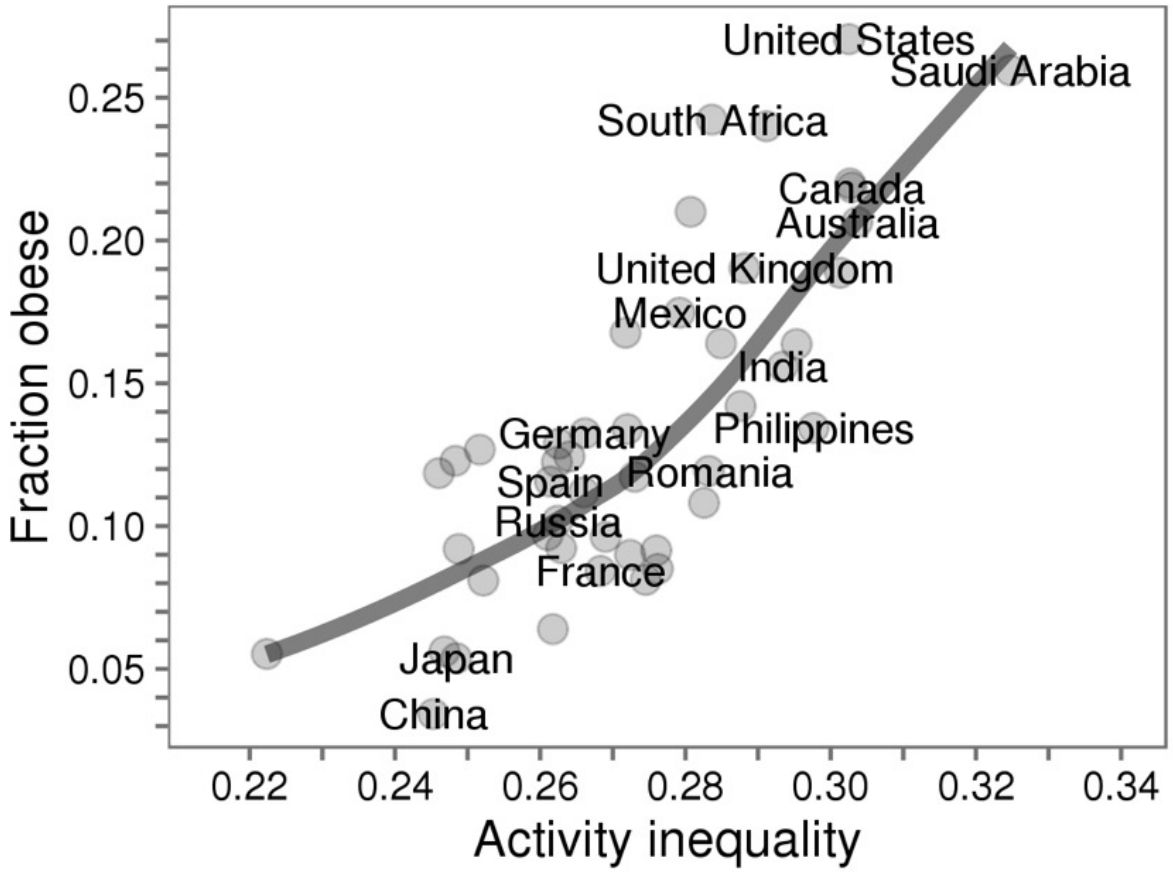
Difference in means



- **How (un)evenly is activity distributed?**
- Gini index of the activity distribution:
  - Activity rich vs. activity poor people

$$G = \frac{\sum_{i=1}^{n}\sum_{j=1}^{n}|x_i - x_j|}{2\sum_{i=1}^{n}\sum_{j=1}^{n}x_j}$$

Result 2: Activity Inequality Predicts Obesity

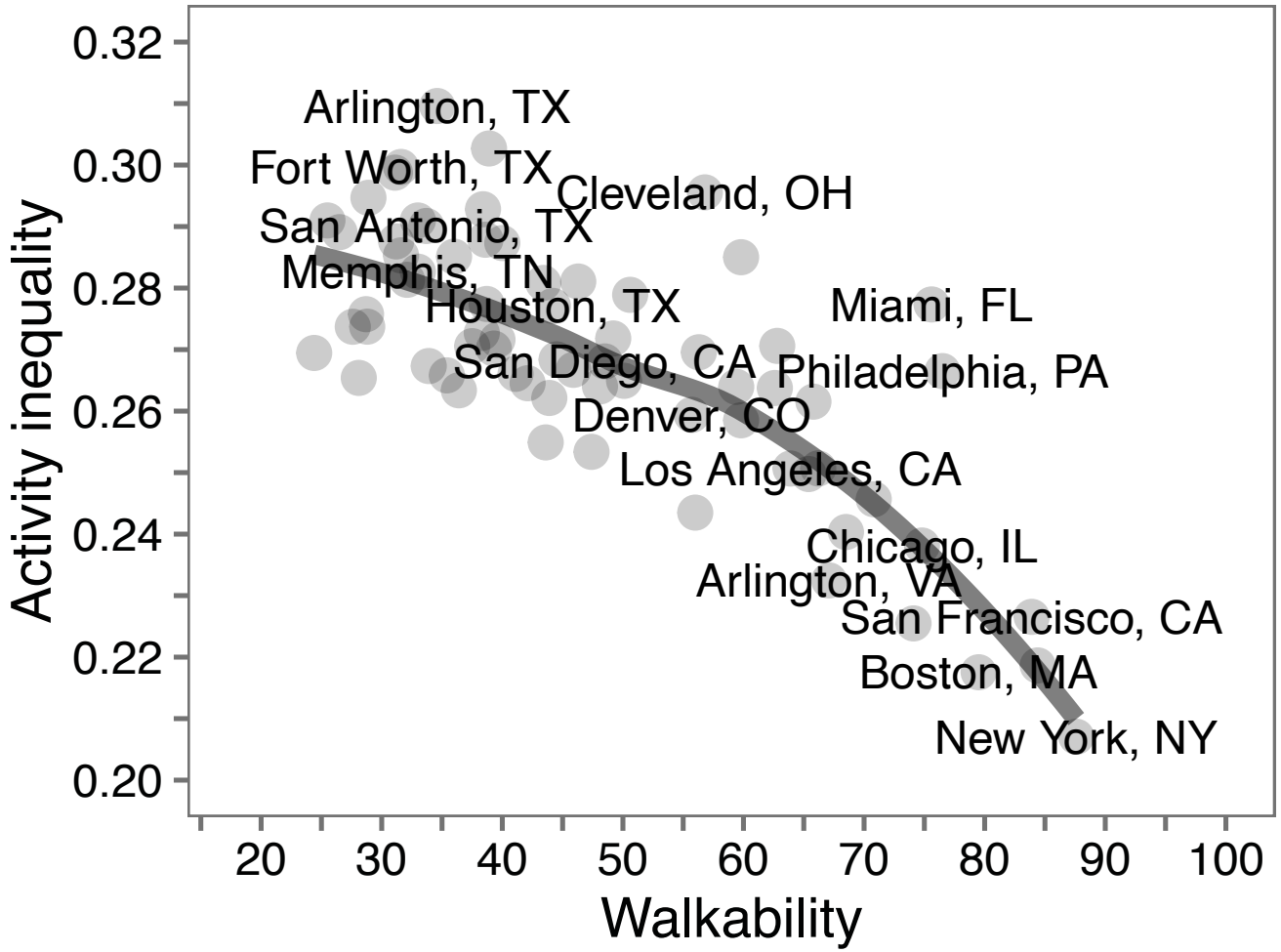


Tails/extremes matter more than the mean

R^2=0.64 (vs. 0.47 for avg. activity)

Massive digital traces uniquely enable studying tails!

Result 3: Walkability Reduces Inequality



Stage 1: Define

Stage 1: Define

- Define the goal and type of the analysis.
- Failure modes: Goal of analysis does not match scientific or business need.

Motivation: How Physically Active Are We?

Physical activity is extremely important for health [Lee et al., 2012]. But we do not know how much physical activity people get!

According to WHO:

- 5-54% of Germans don't get enough activity
- No data for Switzerland and Israel

Health research limitations today:

- High cost, short-term, limited scale
- Biases from self-reporting

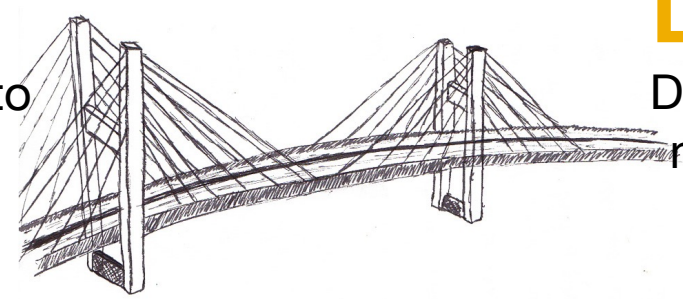
Concrete Research Questions

1. How does physical activity vary across and within countries?
2. What are the relationships between physical activity disparities, health outcomes such as obesity, and modifiable factors such as the built environment?

How to figure out the right question? Engage domain experts!

Data Experts

Don't know what questions to ask & scientific impact



Domain Experts

Don't know data and how new methods could address their big questions

Gaining insights requires intersection of

- Knowing CS methods to extract insights from massive data
- Knowing data, its limitations, and how to address them
- Knowing big questions and how to find new ways to address them

Stage 2: Collect

Stage 2: Collect

- Measure / collect data to analyze.
- Failure modes: Selection bias (e.g., population mismatch, selective labeling…).

In our study…

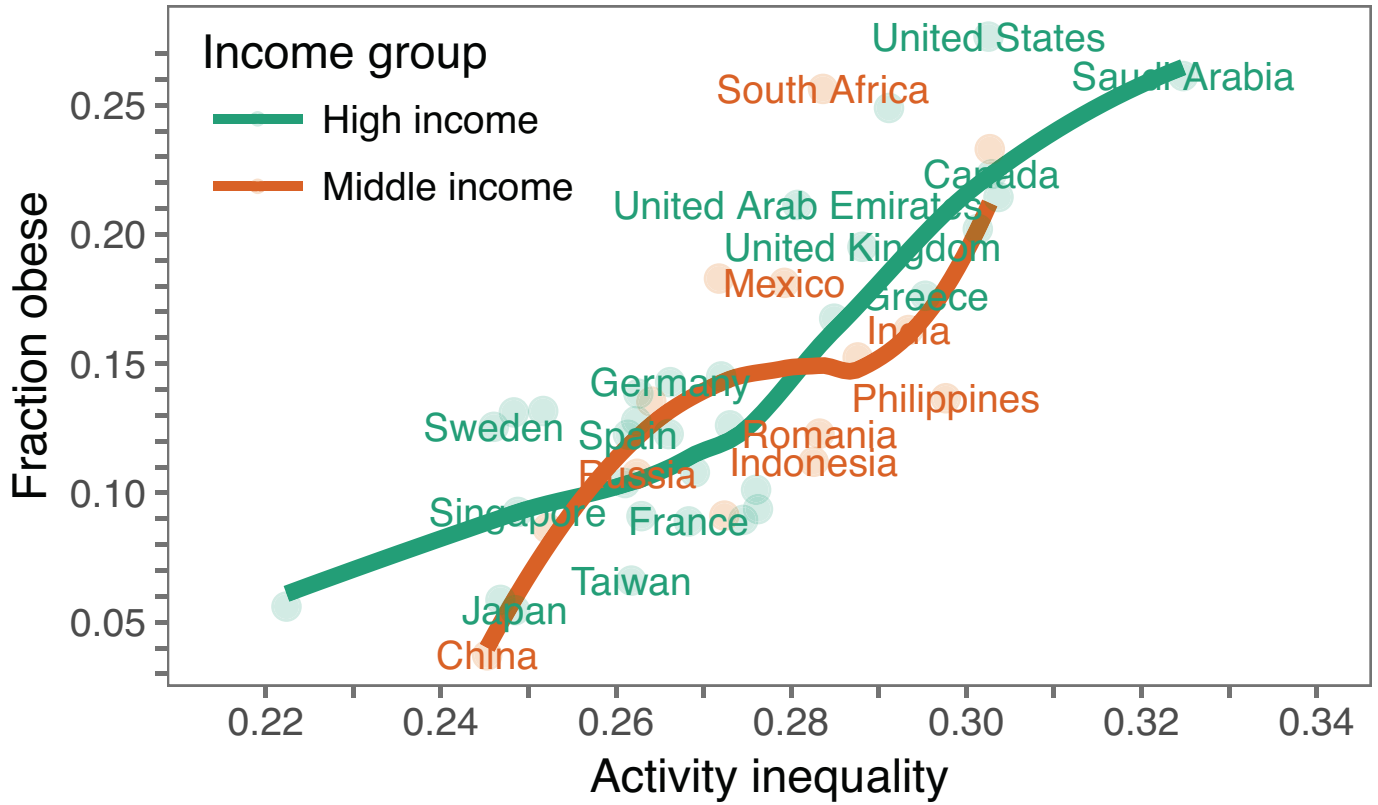
- Key idea was to use smartphones people already have and already collected data – a “convenience sample”
- Selection bias is likely: Participants need to own smartphone (rich?) and use a particular health app (more interested in their own health?)
  - Which population do we want to reason about?
  - Need to investigate selection effects

Who Is Using These Apps?

- Demographics fairly representative in US (but not everywhere!)
  - Gender: 50.2% female (publ.: 50.8%)
  - Age: median 35 years (publ.: 37 years)
  - Obesity: 28.8% (published: 29-35%)

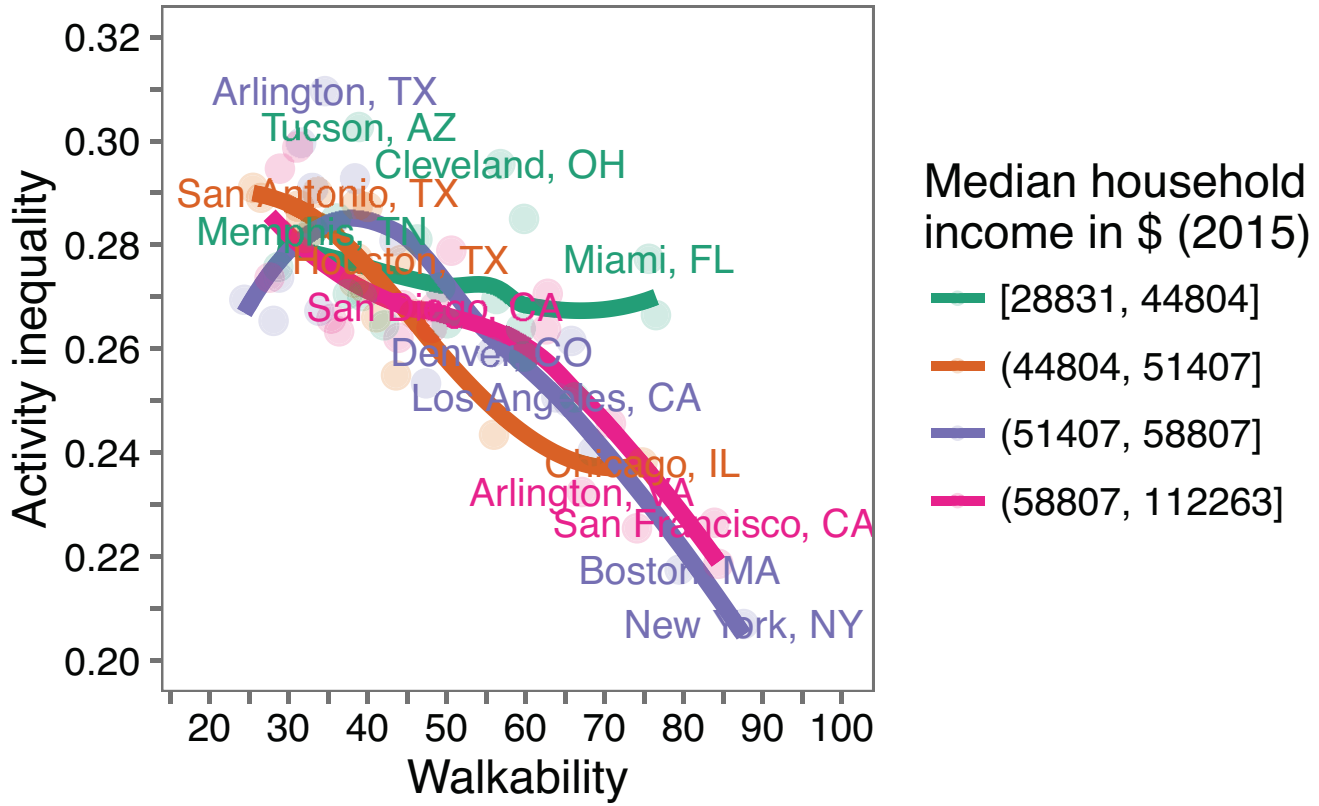
- Method: Resample data according to representative estimates
  - Majority of public health studies may not be representative either – scale allows adjusting

What About Socioeconomic Factors?



- Association between activity inequality and obesity holds within country income groups

Income on City Level



- Relationship between walkability and activity inequality holds within cities of similar median household income (quartiles)

Selective labeling? Is your data missing data at random?

- Goal: To verify that subjects with missing data on gender, age, or BMI are not different from those who report data.
- Method: We computed the standardized mean difference (SMD) in age, gender, BMI, and average steps per day between groups with and without missing data.
  - SMD: Difference in means between groups divided by the standard deviation among the groups

$$\beta = \frac{\mu_1 - \mu_2}{\sqrt{\sigma_1^2 + \sigma_2^2}}.$$

- Result: "Across all combinations of missing variables (age, gender, BMI) and outcomes (age, gender, BMI, daily steps), the groups were balanced, with all standard mean differences lower than 0.25."
  - Implication: The selective labeling of age, gender, BMI was *not awfully biased*

Stage 3: Annotate

Stage 3: Annotate

- Augment data with labels or other metadata.
- Failure modes: Annotator disagreement; erroneous codes or labels.

No Location Data!

- **Game over?** Data did not include location information (e.g. GPS)

But:

- **Phones sync regularly with server**
  - Can get IP address from requests and localize
  - Works well on country level (but not city level) (>99% accuracy)
- **Newer app versions get weather updates and use nearest cell tower**
  - Can get city level location from cell tower (>99% accuracy)

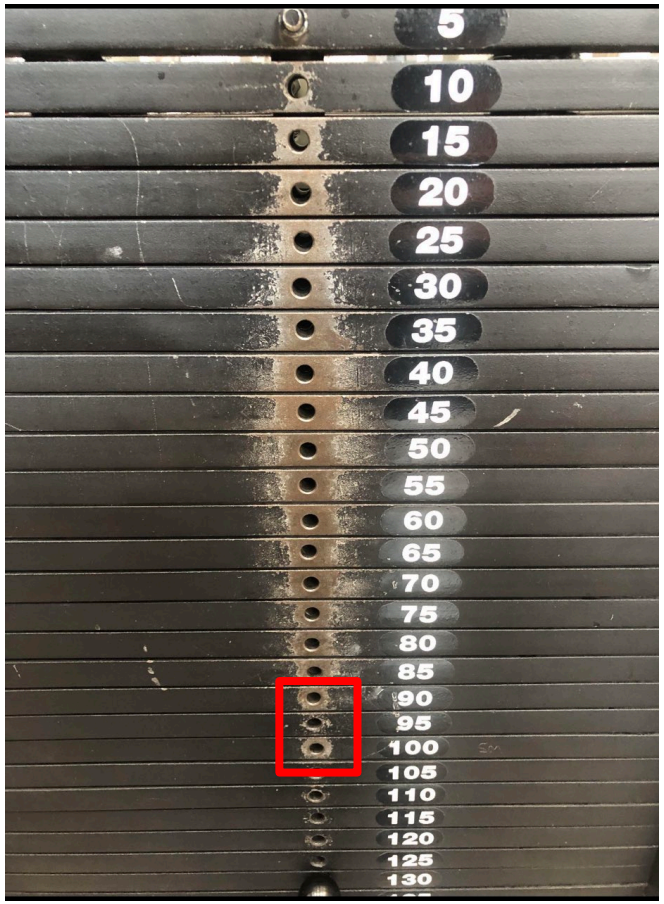
Can we trust self-reported height and weight?

- Consider self-reported height and weight as annotation here
  - Also makes sense to ask these questions when considering construct validity for obesity measures (we'll come back to this)

- Lots of studies have shown that people generally overreport their height and underreport their weight (say ~1lbs).
  - Essentially we like rounding up our height and round down our weight.
- Often this effect is systematic but minor.
  - Some studies have reported 0.99 correlation between self-report and actual weight.
  - We still have to ask ourselves these questions and investigate!

Behavioral Biases: Example

- Distribution is “approximately normal”, but 95 always rounds to 100.



Stage 4: Wrangle

Stage 4: Wrangle

- Clean, filter, summarize, and/or integrate data.
- Failure modes: Incorrect filtering, e.g., high-leverage outliers. Incorrect joins with other datasets.

Basics

- To avoid incorrect filtering or incorrect joins always check your output size and output examples
- These issues are bugs and tend to be simpler than “real validity issues”, which look fine from a join and output perspective but the exact values are systematically biased in some way

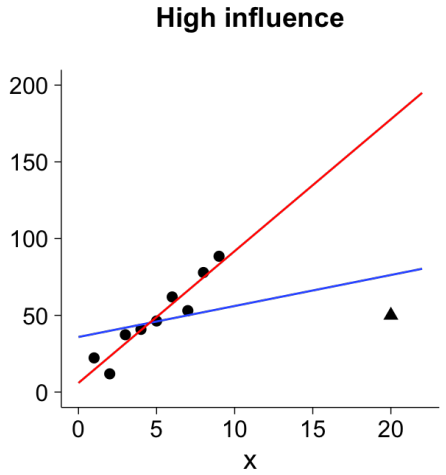
- This stage often takes a lot of time, but often are reported only in minimal ways.

Data Filtering in the Paper

- Ignore days with total steps under 500 (didn't move at all) or over 50k (ran more than a marathon)
  - Filtering criteria based on prior work or domain knowledge
- How many data points do you need to get a reasonably representative picture?
  - You can sample and check the variance in your estimate
  - You can use external data to validate (e.g. correlate with existing PA data, see later)
- Final filtering in this study:
  - Per user: At least 10 days of tracking (average around 90 days)
  - Per country: At least 1000 users
- Always check whether your key results depend on these "harmless" parameters. You'll be surprised!

Can your effect be explained through outliers?

- Sometimes very few data points can have extraordinary impact on the result.



- Paper: ”We further verified that the relationship between activity inequality and obesity is not unduly driven by outliers. We removed the potential outliers of Indonesia, Malaysia and the Philippines from our dataset and found that activity inequality was still a better predictor of obesity than average volume of steps recorded (R^2 was 0.69 for activity inequality versus 0.56 for average steps).”

- Takeaway: Always check for outliers, especially in smaller datasets, or depending on your loss function (squared difference for regression above)

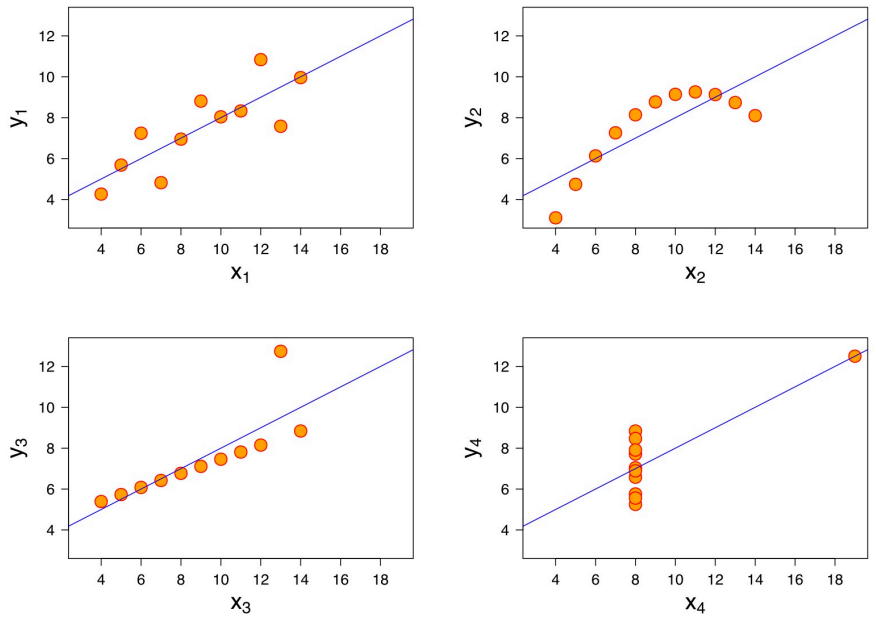
Stage 5: Profile

Stage 5: Profile

- Inspect shape and structure of data.
- Failure modes: Overlook data quality issues or violations of distributional assumptions.

Why inspect your data?

- Always “stay close” to your data and question it!
- Anscombe's quartet: All datasets have identical mean and variance of x and y, similar correlations and regression lines



No real dataset is ever "clean"…

- Profiling allows you to check for errors
- For example, we found:
  - Negative age or weight
  - Age >> 100
  - Daily steps >> multiple marathons
  - IP geolocation does not work well on city level (e.g. all mobile phones get same IP range for some carriers)

- Danger of never-ending pursuit of a "clean" dataset.
  - Do your best in profiling the data.
  - Choose a pragmatic approach: Instead of perfect dataset, can you demonstrate that your finding is robust, even if certain records were changed?

- To avoid distributional assumptions we use Bootstrapping
  - Bootstrapping assigns measures of accuracy (bias, variance, confidence intervals, ...) to sample estimates. This technique allows estimation of the sampling distribution of almost any statistic using random sampling methods (with replacement).
  - We don't need to make assumptions like "X follows a Normal distribution"

Stage 6: Operationalize

Stage 6: Operationalize

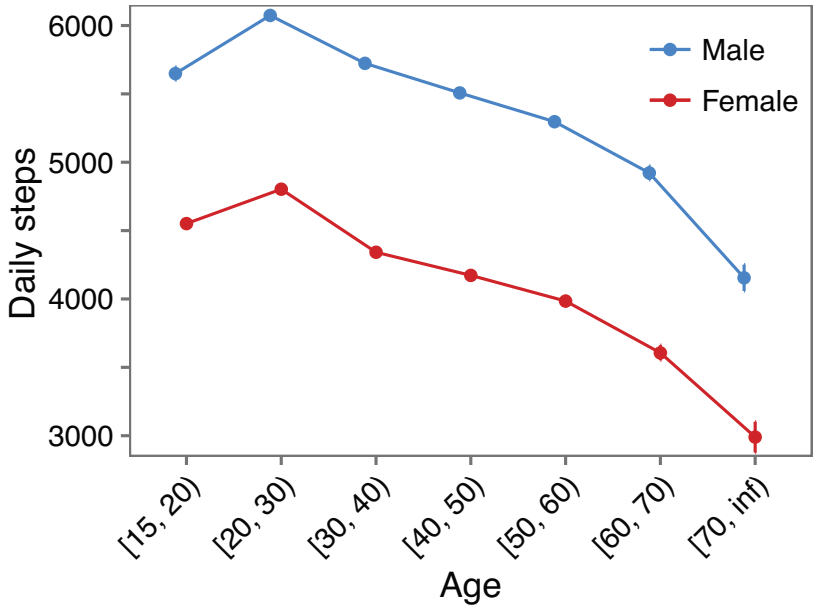
- Define and validate central measures, which may be proxies. Failure modes: Lack of construct validity (i.e., not measuring what you think you are measuring).

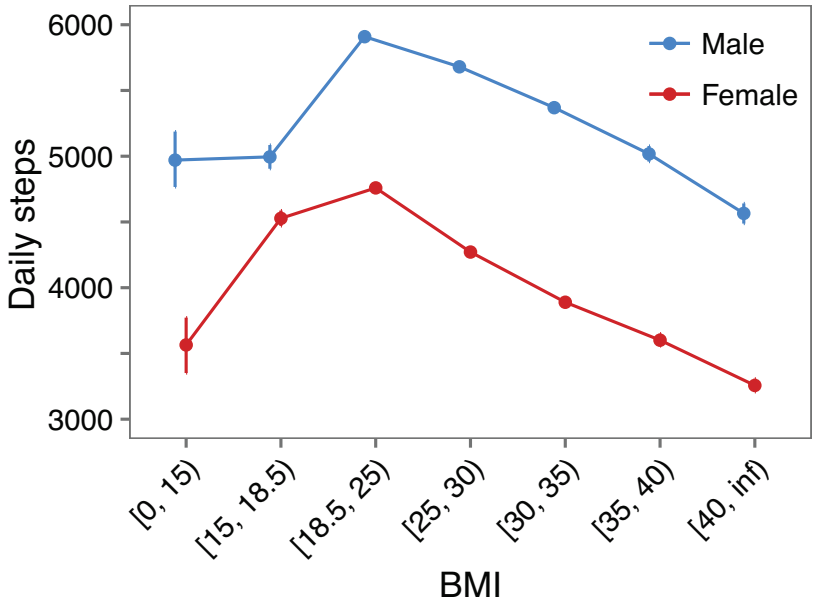
In our case…

- Is **physical activity** appropriately measured by smartphone-determined steps? Does wear time add a systematic bias?
- Is self-reported **obesity** data trustworthy?
- Should **activity inequality** be operationalized through the Gini coefficient or some other measure of inequality?
- Does WalkScore appropriately quantify the **walkability** of the environment?

- **This step is absolutely critical and important!**
  (and too often ignored or taken lightly)

Validity of New Sensor: Smartphone Steps ~ PA?

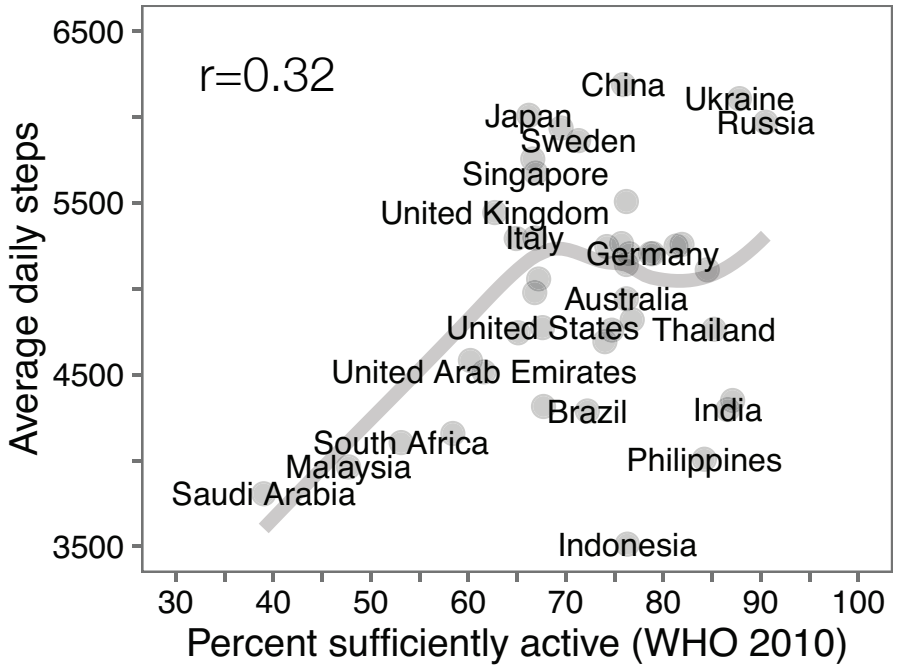


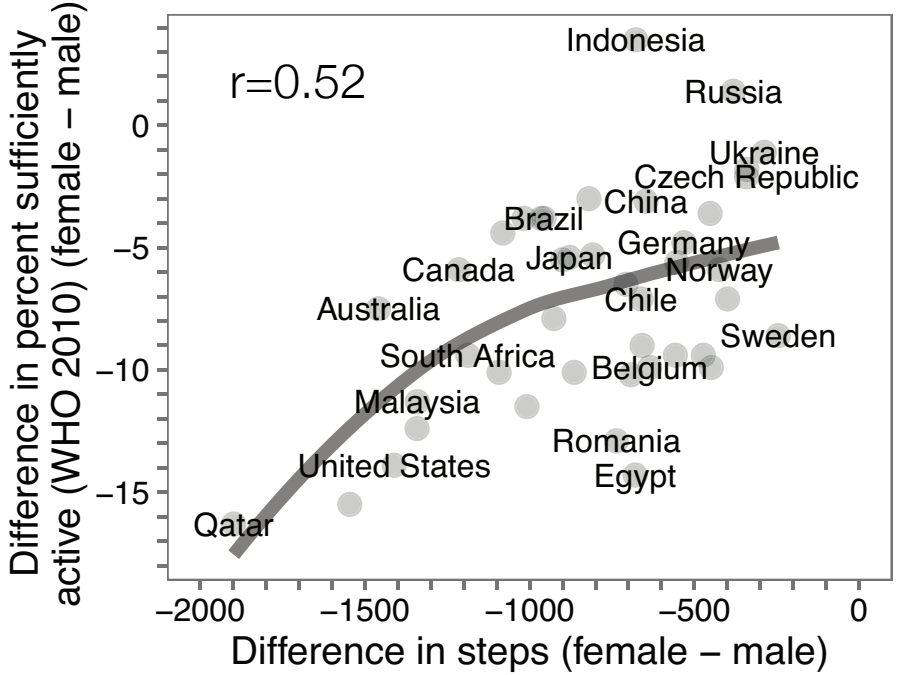


- Smartphone data reproduce established relationships between age, gender, weight status, and activity
- Smartphones provide accurate step counts [Case et al., 2015] [Hekler et al., 2015]

[Hallal et al., Lancet 2012]
[Bauman et al., Lancet 2012]
[Van Duck et al., Int. J. Obes. 2015]
[Bassett et al., MSSE. 2010]
[Troiano et al., MSSE. 2008]
[Tudor-Locke et al., MSSE. 2009]

Prior Physical Activity Data



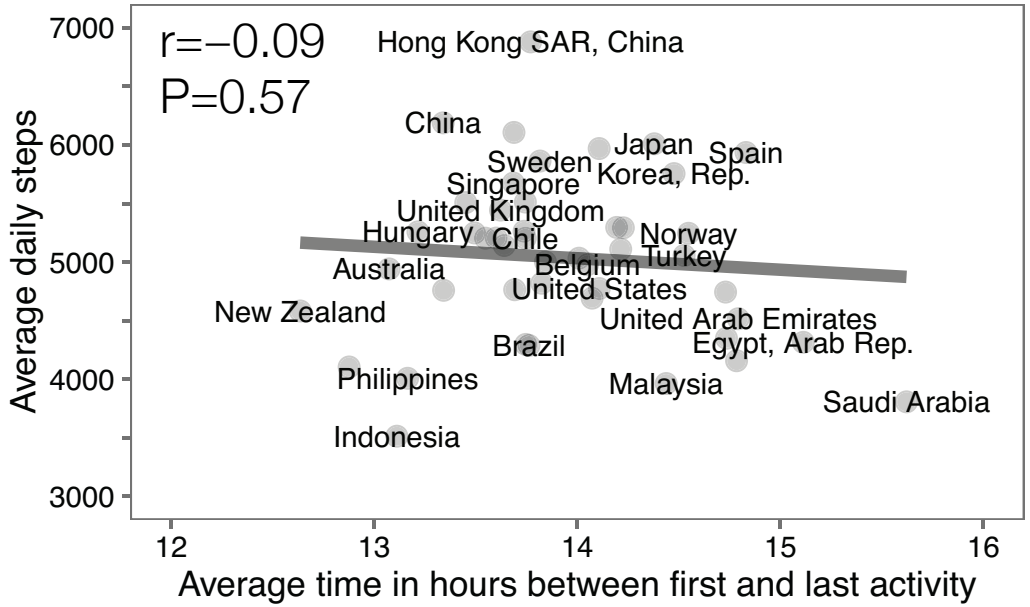


Limitations in comparison:

- % active vs. avg. daily steps
- Self-report vs. accelerometry can differ [Prince et al., 2008]
- WHO confidence intervals are very large

How Does Wear Time Affect Results?

- Daily span of recorded activity (time between first and last recorded step each day)



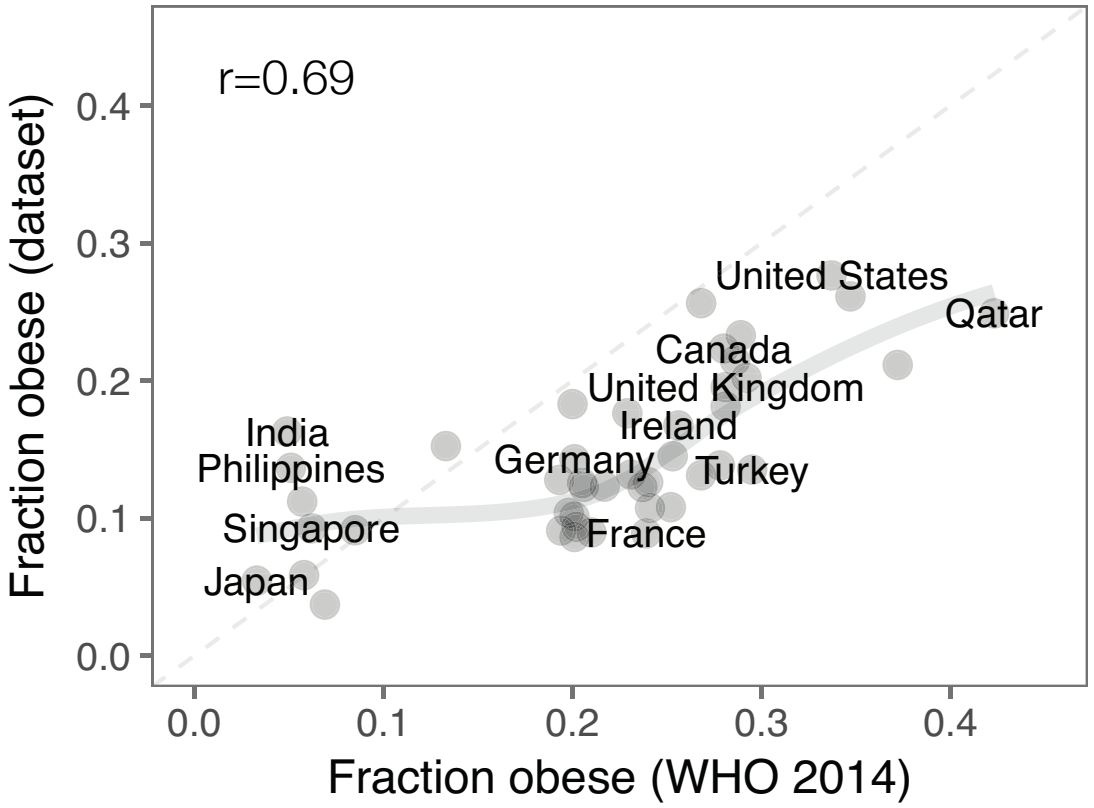
- What about disparate impact on women?

  30min shorter span compared to men

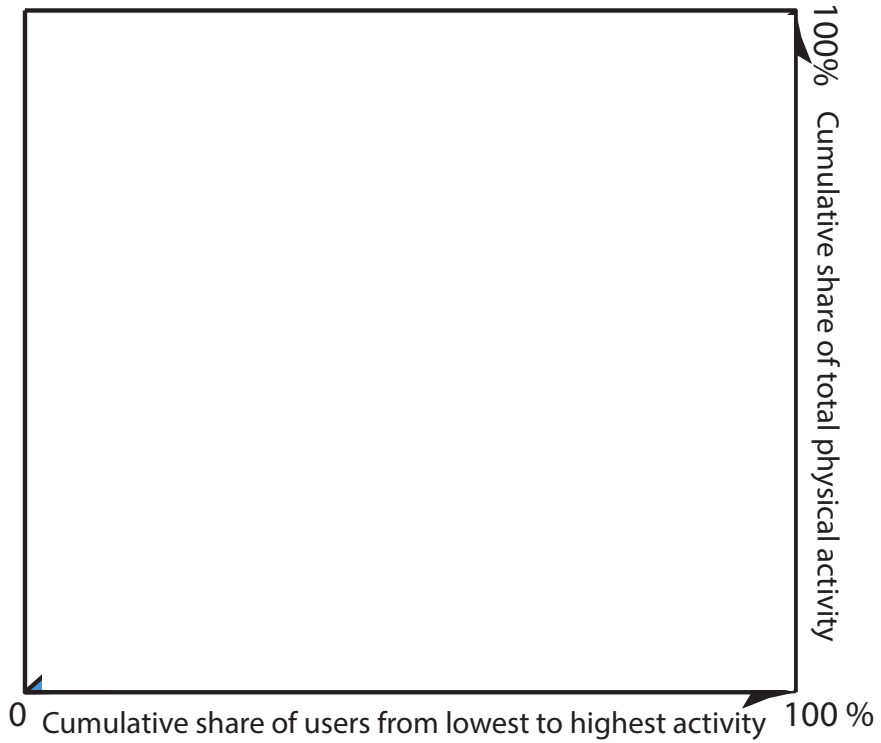
Do Women Wear Phones Less?

- Yes, women 30min shorter average wear time than men (13.7h vs. 14.2h)
- Consistent with longer average sleep duration of females
  [Walch et al., Sci. Adv. 2016] [Basner et al., Sleep 2007]
- Similar gender activity gap to previous studies using both accelerometry and survey measures

Prior Surveillance Data: Obesity



Choice of inequality measure



- Activity Inequality: Gini coefficient of activity distribution

What about other ways to quantify inequality?

- "Several other measures have been used to quantify inequality and statistical dispersion including the coefficient of variation, decile ratio, and others; we find that these measures are all highly correlated with the Gini coefficient (r=0.96 or higher) when applied to step counts within countries."

→ It doesn't matter. They all are virtually identical in this case.
→ But again we still need to investigate!

Operationalizing walkability

- There is no agreement in the literature on the best measure.
  - Without large-scale objective measurements it's impossible to evaluate which measure is better.
- Used WalkScore.com because it was freely available and used in prior literature. Our domain scientist collaborators seemed to trust it.

City	Zip Code	Walk Score	Transit Score	Bike Score	Population
Seattle *(the largest city in Washington)*	98115	74	60	70	608,660
Tacoma	98406	53	--	49	198,397
Lynnwood		53	--	50	35,836
Burien		51	39	46	33,313
Yakima	98902	50	27	50	91,067
Mountlake Terrace	98043	50	--	43	19,909
Spokane	99205	49	36	52	208,916
Bellingham	98225	49	37	58	80,885
Bremerton	98337	49	31	38	37,729
Everett	98203	48	39	55	103,019
Shoreline		48	43	50	53,007
Kirkland	98033	48	40	50	48,787
Wenatchee		47	33	47	31,925
Walla Walla		47	25	63	31,731

Research Using Walk Score Data

- Pivo, Gary, and Xudong, An. 2016. Sustainable Development and Commercial Real Estate Financing: Evidence from CMBS Loans.
- Christopher B. Leinberger. 2015. Why American Companies are Moving Downtown. The George Washington University School of Business.
- John I Gilderbloom, William W. Riggs, and Wesley L. Meares. 2015. Does walkability matter? An examination of walkability's impact on housing values, foreclosures and crime. Cities, Volume 42, Part A.
- Christopher B. Leinberger. 2015. The WalkUP Wake-Up Call: Michigan Metros. The George Washington University School of Business.
- Christopher B. Leinberger. 2015. The WalkUP Wake-Up Call: Boston. The George Washington University School of Business.
- Ethan N. Elkind, Michelle Chan, Tuong-Vi Faber. 2015. Grading California's Rail Transit Station Areas. Center for Law, Energy and Environment, University of California, Berkeley.
- Pivo, Gary, and Xudong, An. 2015. Default Risk of Securitzed Commercial Mortgages: Do Sustainability Property Features Matter?
- Christopher B. Leinberger. 2014. Foot Traffic Ahead. The George Washington University School of Business.
- Christopher B. Leinberger. 2013. The WalkUP Wake-Up Call: Atlanta. The George Washington University School of Business.

5 min break ☺

Cluster into your groups

Your turn ☺ [10 min in group, 10 min report]

- Phase 6: Operationalize
  1. What are the central constructs in your research questions?
  2. How will you operationalize them? What would be alternative ways of doing so? How does domain knowledge inform your decision?
  3. Concretely, how can you validate your choices (e.g. how to demonstrate discriminant and convergent validity)?

- After group discussion – report back for 1-2 minutes
- Take notes for your validity reflection assignment

Stage 7: Explore

Stage 7: Explore

- Interactively explore data and variable relationships.
- Failure mode: Confirmation bias; unclean split between train/test data.

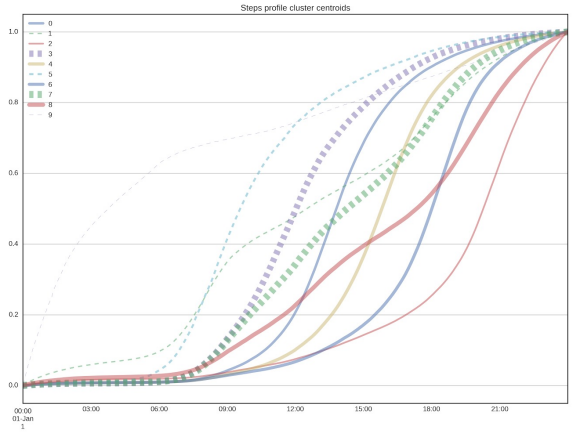
Exploring the data

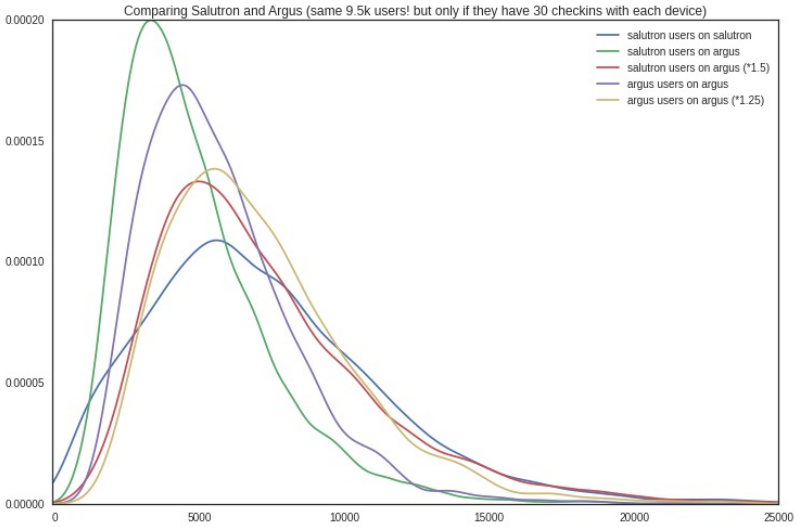
- Not well documented like the research paper but critical
- Created around 10,000 visualizations over multiple years before we arrived at the final publication

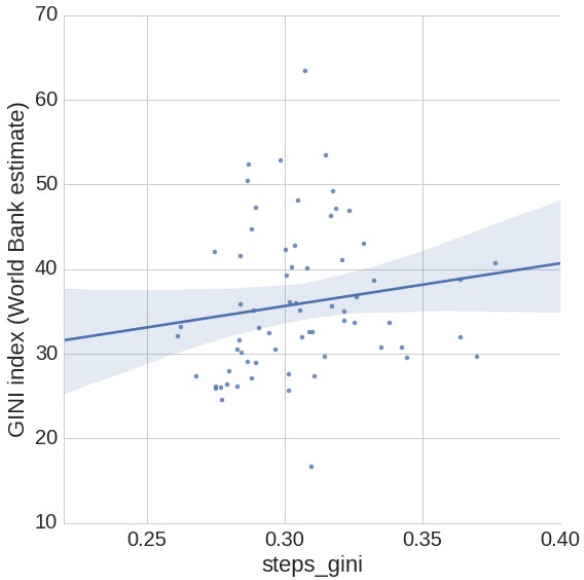
- Always look at the raw data. Investigate. Interrogate your dataset. See where you can trust it. What's broken? Every dataset is messy!

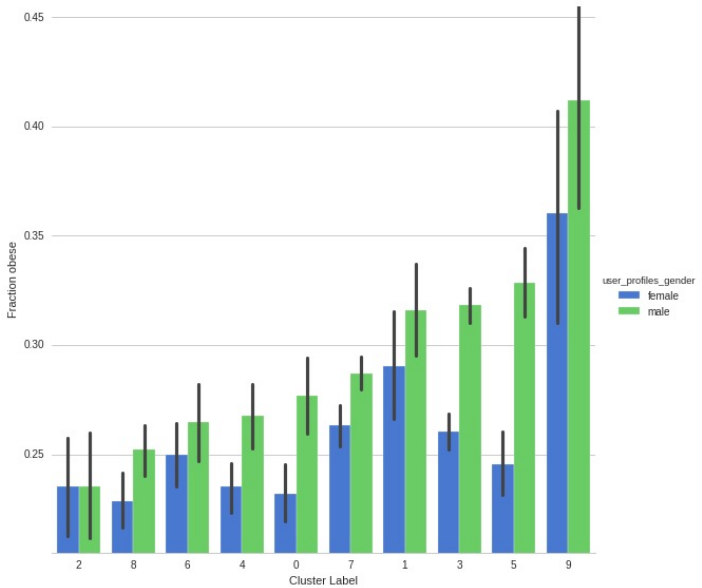
- Do not put lots of layers of models and statistics between you and the data.

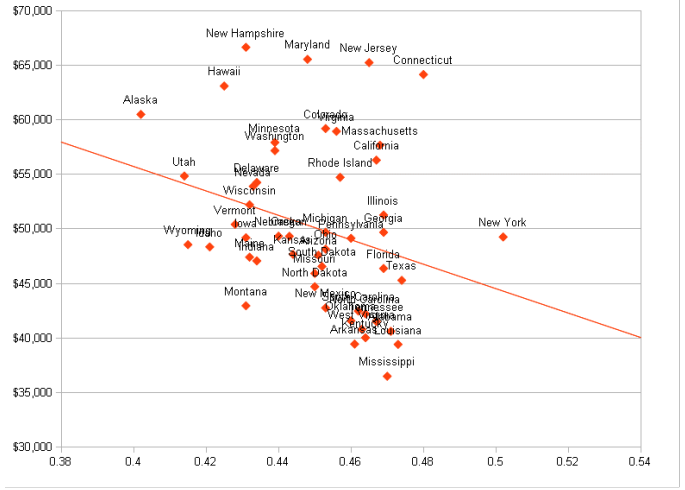
Example explorations along the way

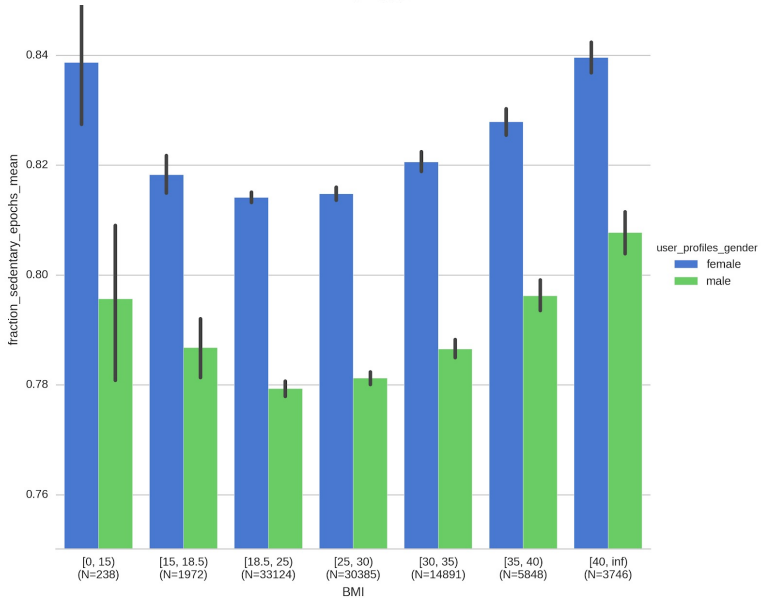




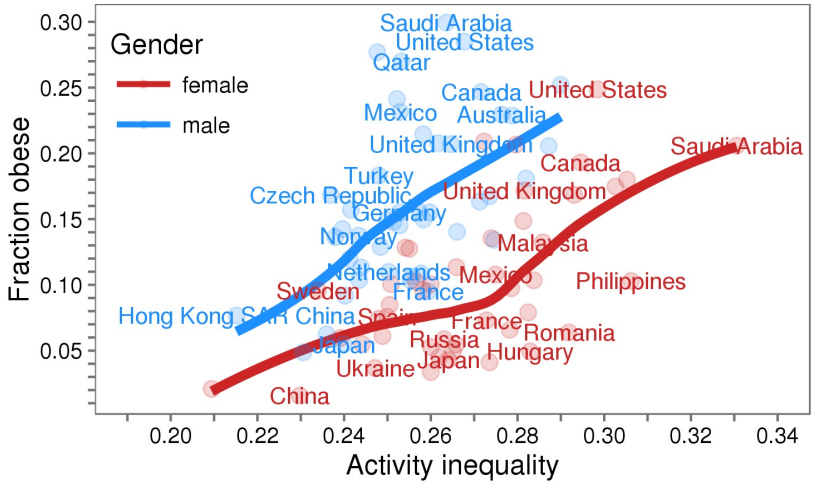


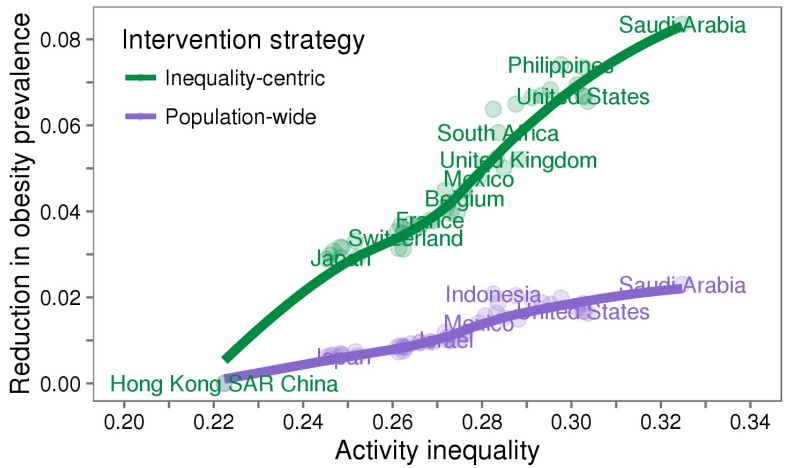


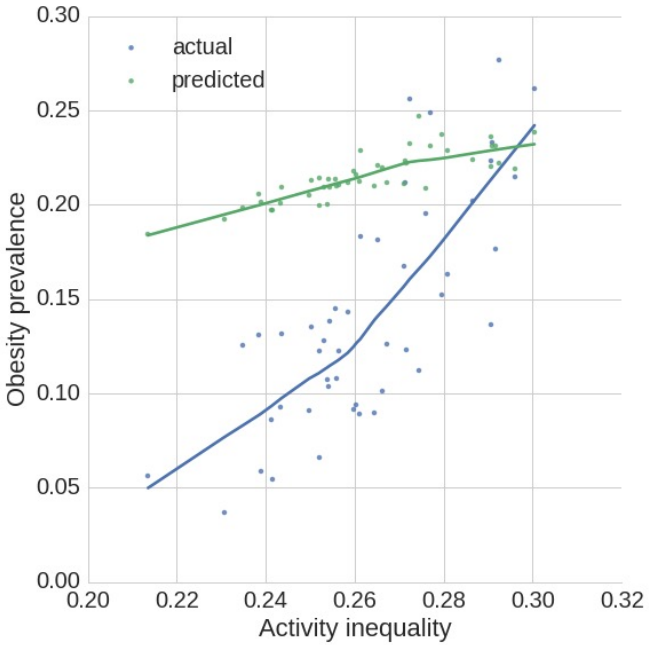


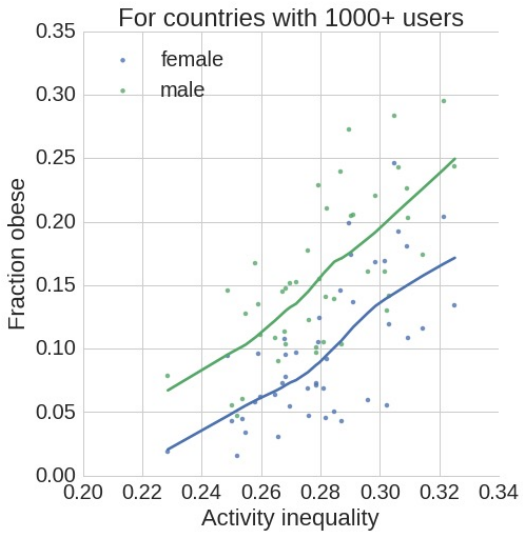


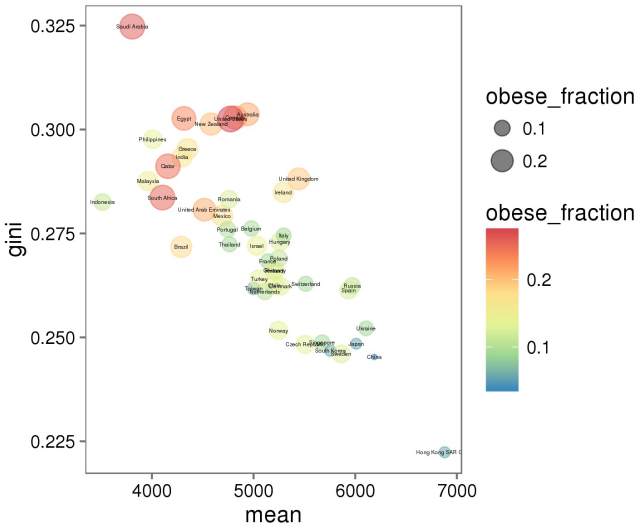
Example explorations along the way (2)











Stage 8: Model

Stage 8: Model

- Define and fit models of relationships in data.
- Failure modes: Lack of internal validity. Failure to identify effect, e.g., due to confounding or violated assumptions.

There is no machine learning model in this paper!

- Data science frequently includes machine learning models
- But it is not necessary!

- Science matters, not your machine learning model!

- Models can also be conceptual: e.g. weight is impacted by both physical activity, diet, and possibly other factors.

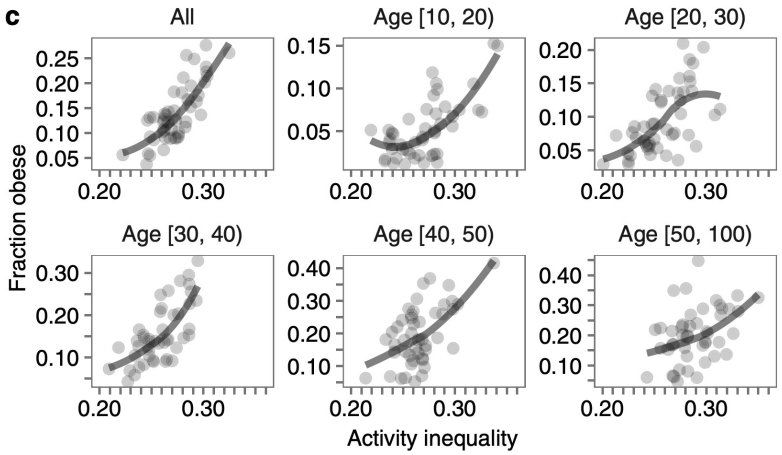
What could be confounding the analysis?

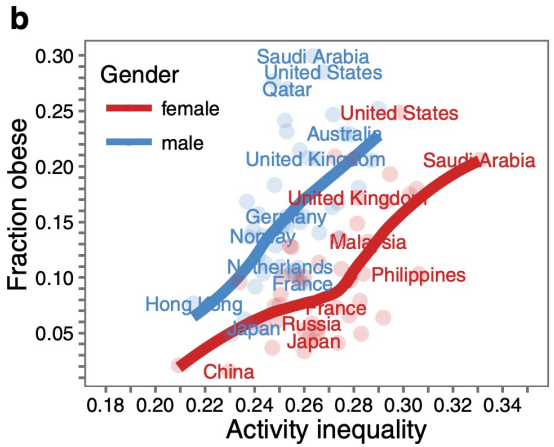
- No random assignment of participants to locations, to activity levels, to tracking, to reporting → we cannot make any causal claims
  - What would be a way to get to better causal evidence?
  - If we made causal claims this would be very poor internal validity.

- Why do we still share share correlational results?
  - Construct validity: PA measures are highly correlated with prior data
  - The results are plausible - yet striking because we couldn't quantify them before
  - The results are actionable – they quantify existing inequality and modifiable factors

Additional analyses to support the main findings

- Sample correction, stratification, outlier, and balance testing
- Conclusions are robust to
  - Missing data (discussed before)
  - Outliers
  - Population bias in age and gender (holding also within these groups)





City-level example: Bay Area

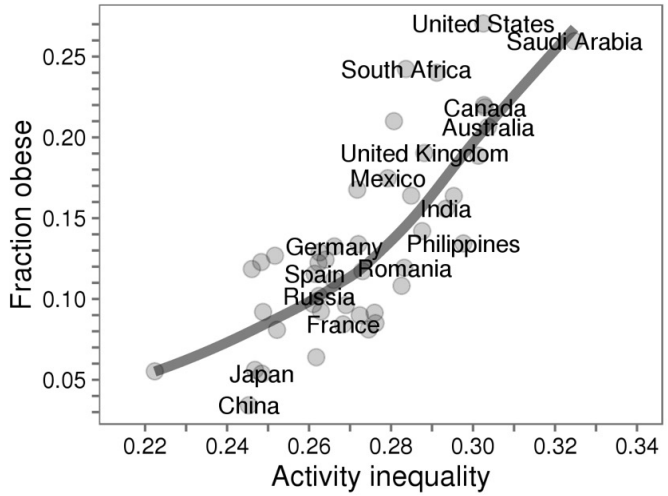
- San Francisco, San Jose, and Fremont
  - Geographically close
  - Socioeconomically similar*

City	Walkability Score	Activity Inequality* (Percentile)
San Francisco, CA	83.9	0.227 (0.07)
San Jose, CA	48.1	0.264 (0.33)
Fremont, CA	44.5	0.268 (0.48)

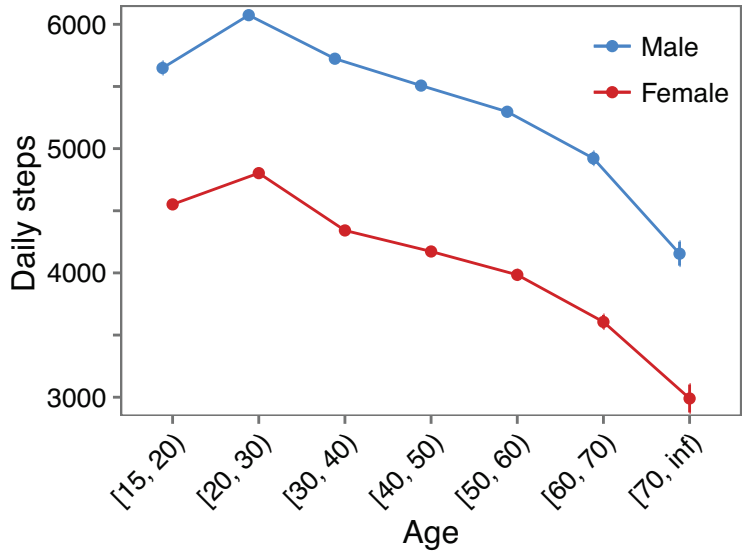
Country-level example: USA vs. Mexico

- Similar daily steps
  - 4,774 versus 4,692
- USA larger activity inequality
  - 0.303 versus 0.279 (10th versus 7th decile)
- USA larger obesity prevalence
  - 27.7% versus 18.1% (10th versus 8th decile)

Model assumptions



- We use LOWESS regression model
  - Very flexible and we still show data points
- Advice: Show your data. Don't put layers and layers of complex models in between you and your data. It will confuse yourself and others!
- Use bootstrapping for non-parametric confidence intervals (no std error etc)



Stage 9: Evaluate

Stage 9: Evaluate

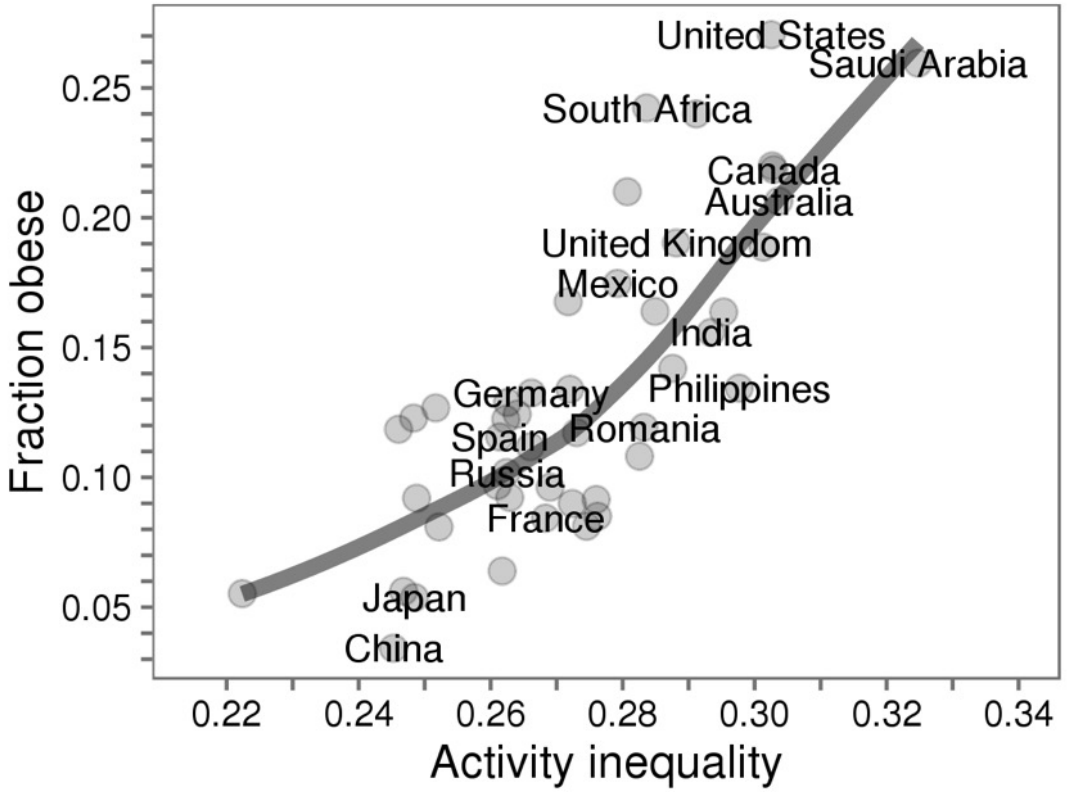
- Measure explanatory power or predictive accuracy of model using appropriate statistical techniques. Failure modes: p-hacking, overuse of test set data.

- Think about how you will tell whether your project was successful? (we will ask you that question!)

Goals in the paper

1. Why should we care? → Demonstrate relevance to health outcomes
2. What contributes to this inequality? → Demonstrate some understanding of the inequality
3. What can we do about it? → Demonstrate a modifiable factor(s) that could plausibly impact inequality

Activity Inequality Predicts Obesity

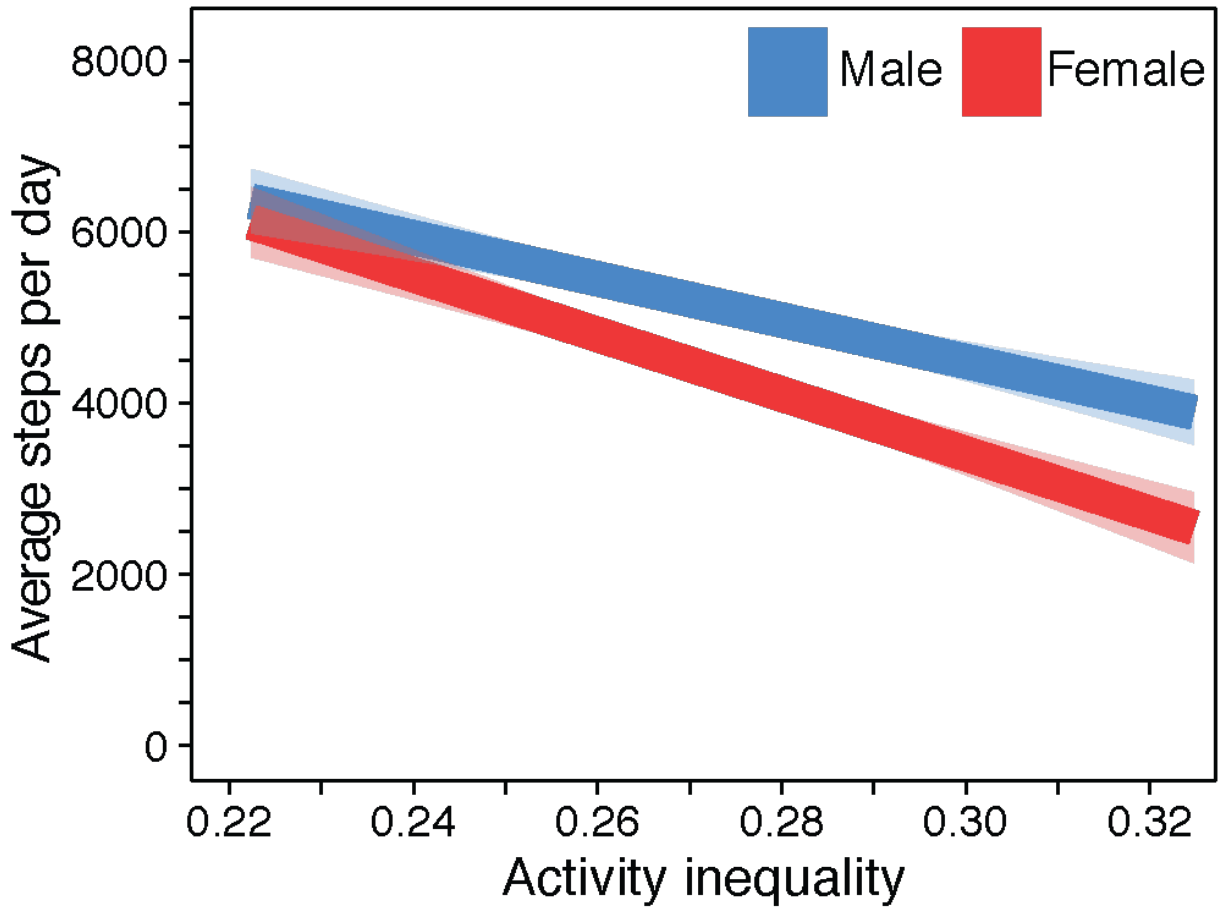


Tails/disparities matter more than the mean

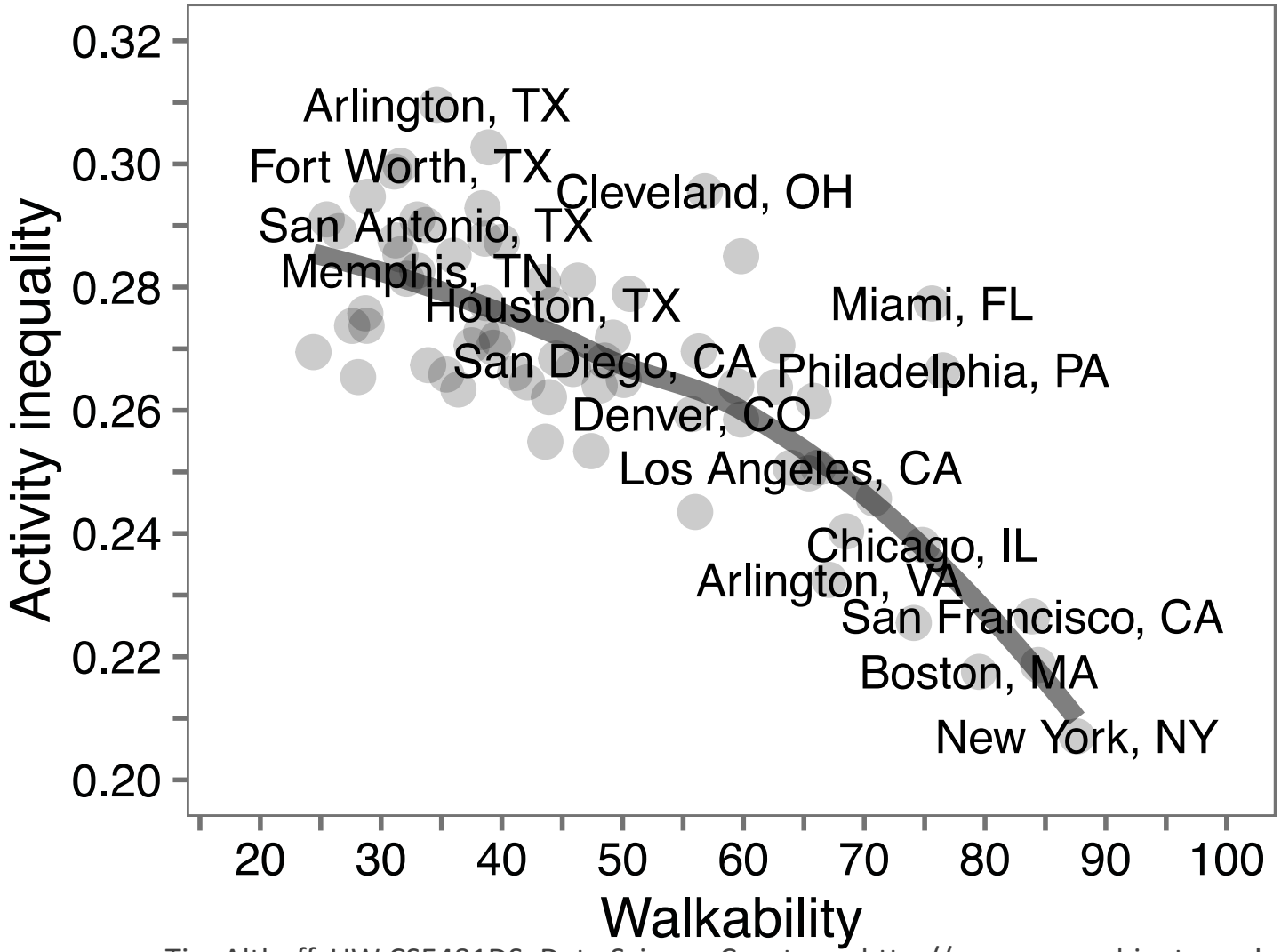
R^2=0.64 (vs. 0.47 for avg. activity)

64% of obesity variance can be explained with a single variable!

Gender Gap: 43% of Inequality



Walkability Reduces Inequality



Stage 10: Report

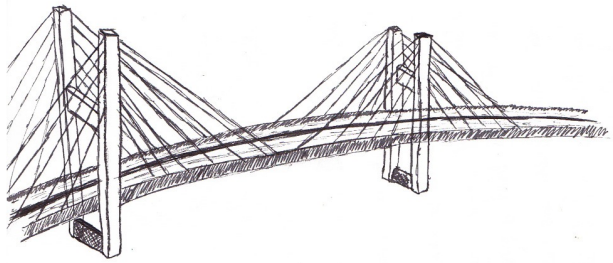
Stage 10: Report

- Report results and potential generalizations.
- Failure modes: Misinterpretation (e.g., generalization, uncertainty), miscommunication via errors or omissions.

- Think about who your audience is? How do you need to communicate your findings? What is the story of your findings?
  - Avoid long "laundry lists" of results. Crystallize your findings into a coherent story. Less is often more!
  - Tell a story! Provide empirical evidence for your story through data.

The Challenge: Convincing Domain Experts

- New construct + new instrument = skepticism
- Domain experts know that these data are ...
  - Noisy
  - Sometimes inaccurate
  - Observational
  - Biased and full of selection effects
- That is why data have been thrown out before
- Designed and conducted over 20 reweighting, resampling, stratification, and simulation experiments to demonstrate validity of results



Demonstrating Validity of Results

…in light of valid concerns

- Flawed sensor?
- But women wear phones less?
- Obesity data inaccurate?
- Biased population?
- Due to rich people?
- Missing data? Outliers?
- Inaccuracy of location inference?
- Reproducible: Publicly released analyses & data

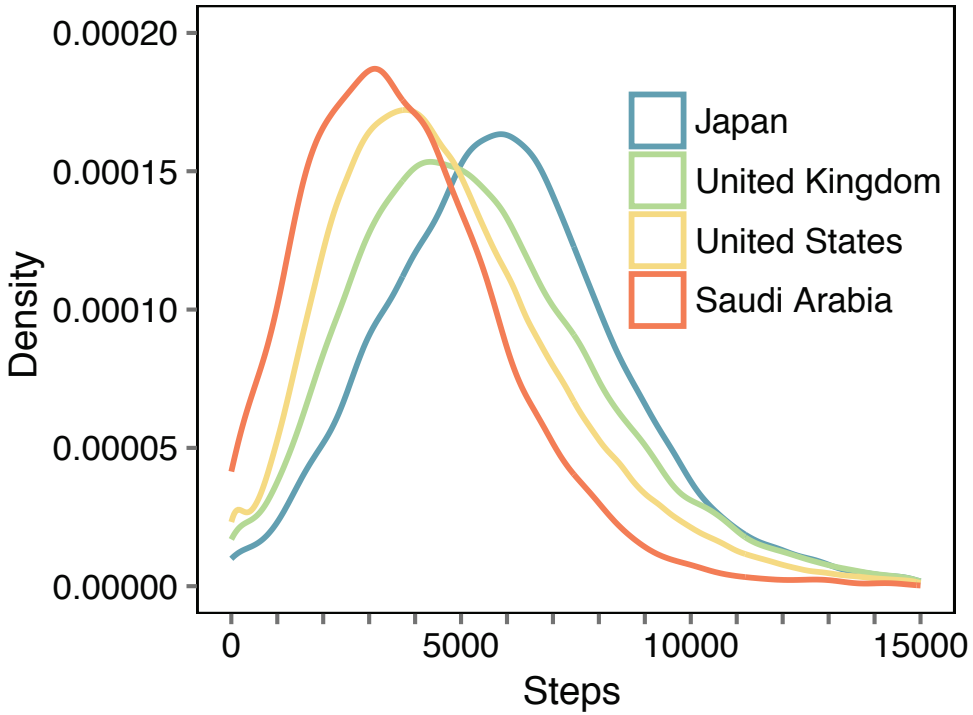
Design your visualizations to effectively communicate your data and results

- We will have a full lecture on data visualization

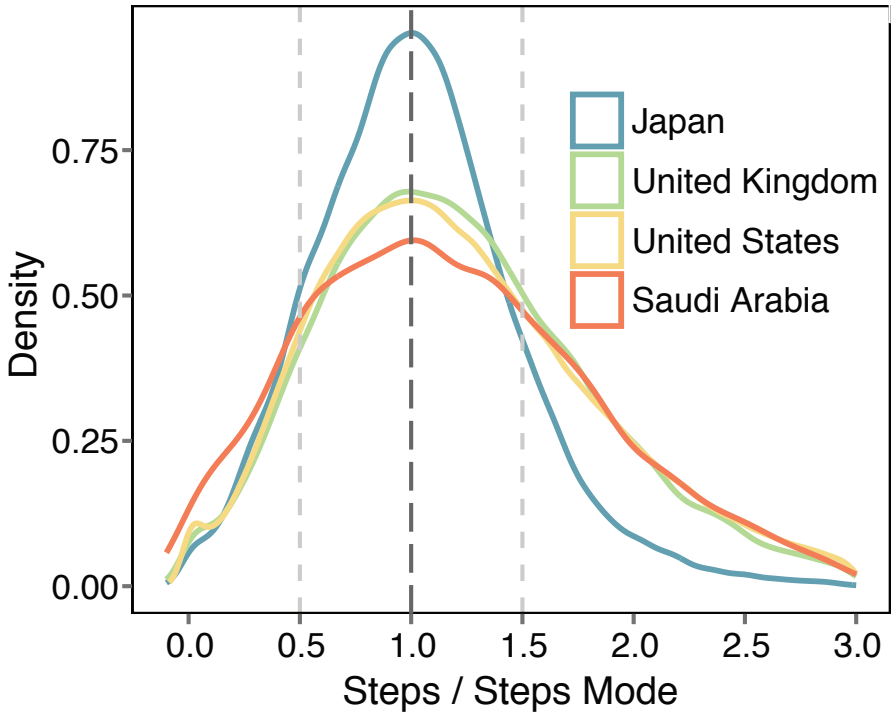
- Show your data – it helps build understanding and trust in your approach

Example 1: Inequality of Physical Activity

Difference in means



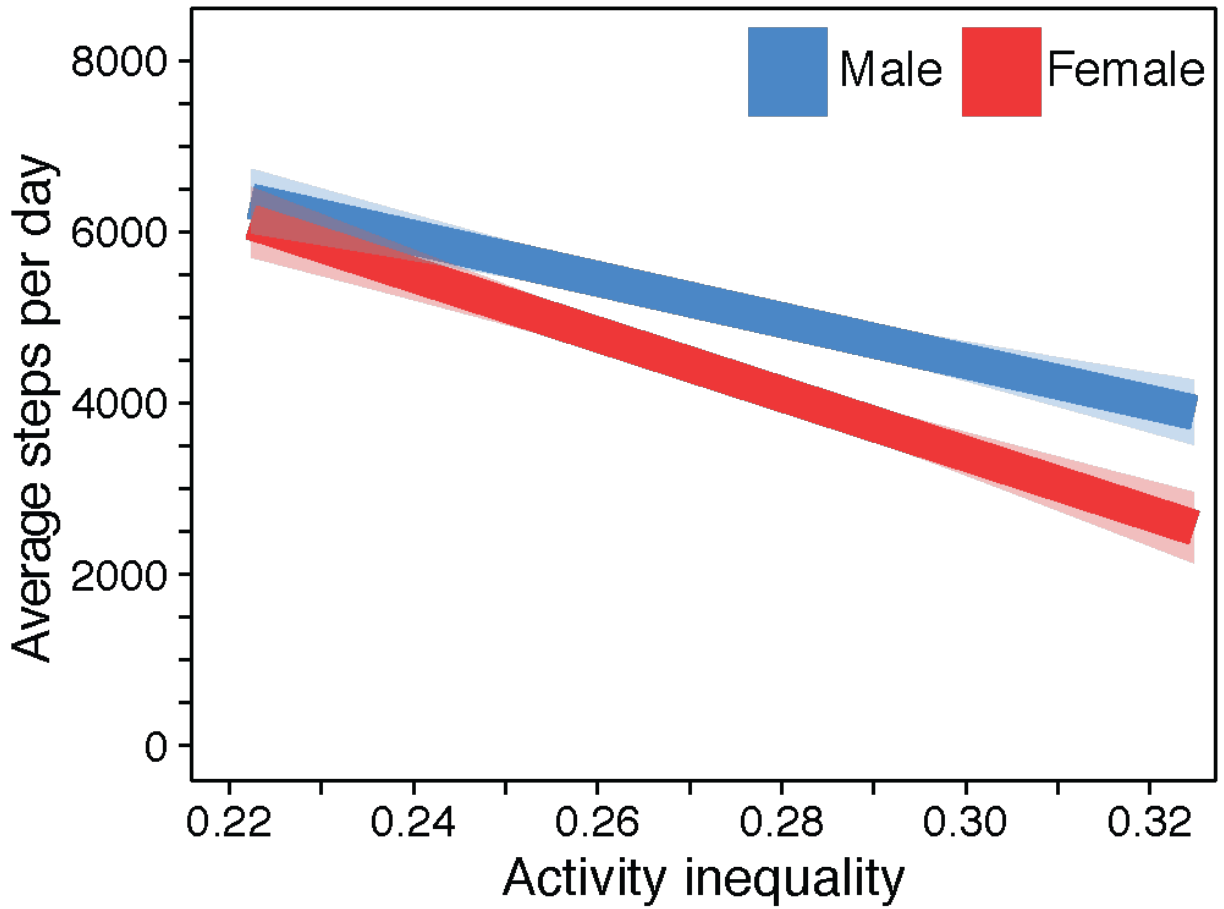
Difference in variance



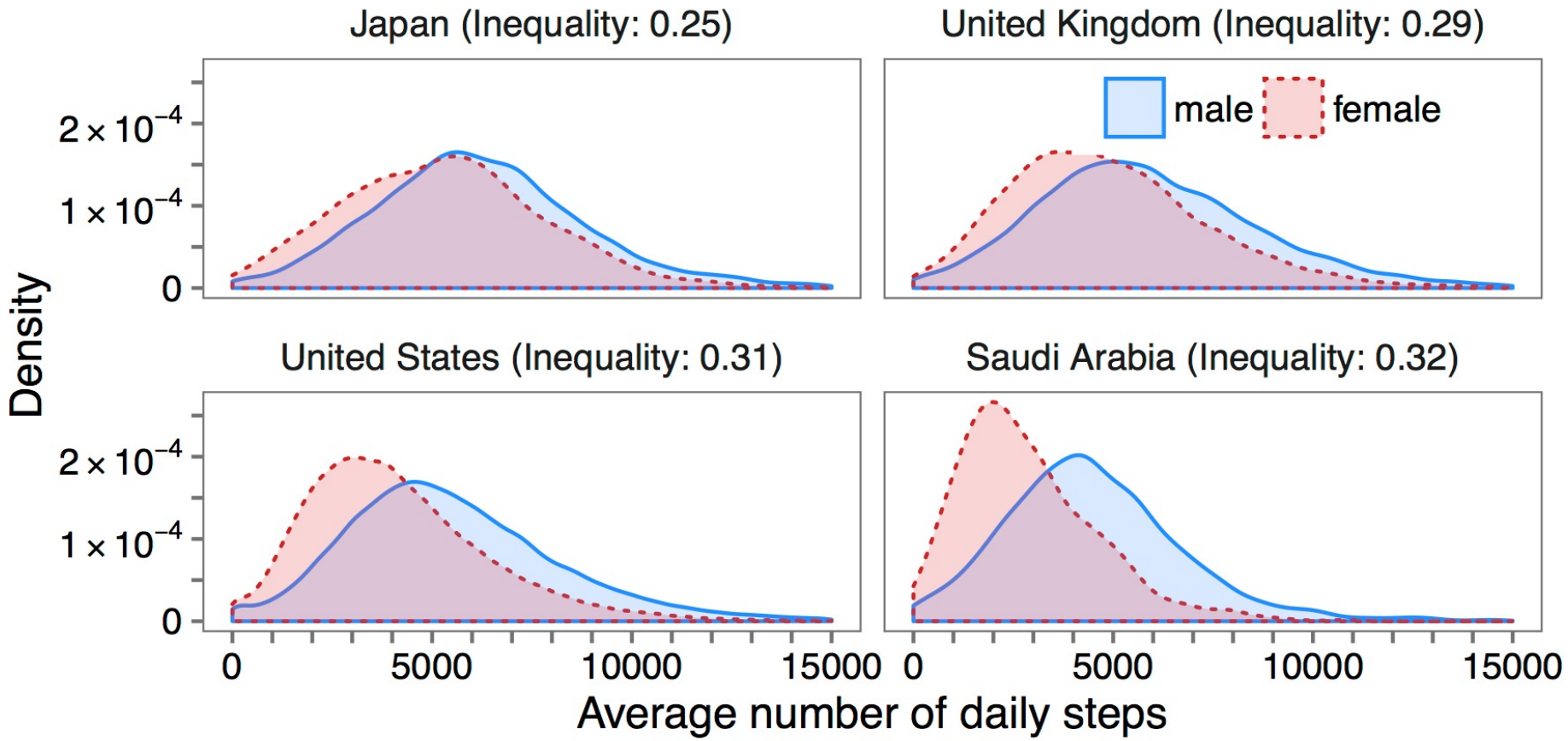
- **How (un)evenly is activity distributed?**
- Gini index of the activity distribution:
  - Activity rich vs. activity poor people

$$G = \frac{\sum_{i=1}^{n}\sum_{j=1}^{n}|x_i - x_j|}{2\sum_{i=1}^{n}\sum_{j=1}^{n}x_j}$$

Gender Gap: 43% of Inequality



Example 2: Gender Activity Gap



Larger variances due to reduced activity in females, not just increase in variance overall!

Stage 11: Deploy

Stage 11: Deploy

- Deploy model or enact decision. Failure modes: Distribution drift, e.g., changes in data pipeline upstream, changing assumptions, adversarial input.

In this case…

- No deployments as part of this research project
- Data product: All data was publicly released for use by others – various cities and countries and journalists used it

- Consider how you can have the largest impact?
- But beware when making decisions based on data, especially when they are automated without humans in the loop.
  - Example: Predictive policing – "optimally" assign resources to areas of high crime. What are the implications?
  - Think about what your projects and products could be used for

- Every study needs a "so what?" or "why should we care?"
  - This won't happen by accident. Design for it from the beginning.

Here: So what? Research Implications

- Pioneered new paradigm for monitoring populations
- Stronger evidence through objective measures
- Publicly released data on physical activity and inequality
- Implications for health policy and urban planning
- Highlighted importance of built environment for physical activity

Recap: Data Science Process

- **Plan** your own project along these stages
- When learning about other projects **pay attention potential pitfalls** across all phases
- When working on your own project, **explicitly address each step and failure modes**

Any Questions?

Thank you for sharing your feedback with us!

https://bit.ly/cse481ds-au22-feedback