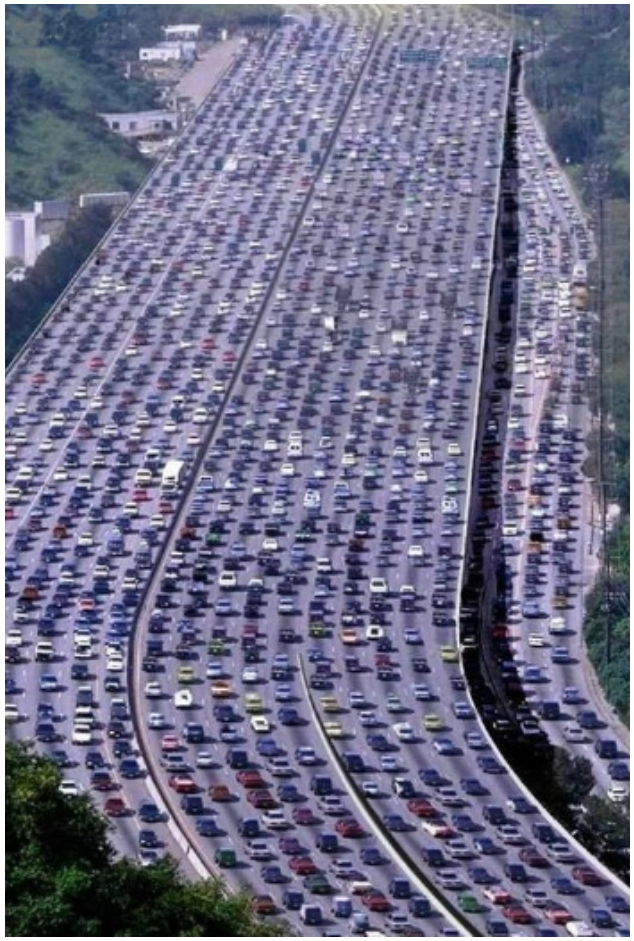
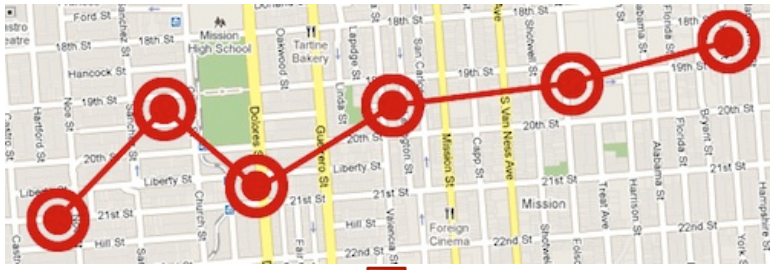


On the Validity of Geo-Social Mobility Traces

Zengbin Zhang, Lin Zhou, Xiaohan Zhao,
Gang Wang, Yu Su, Miriam Metzger,
Haitao Zheng, Ben Y. Zhao

University of California, Santa Barbara
Hotnets, November 2013

Understanding Human Mobility



Traffic Planning

Controlling Infections

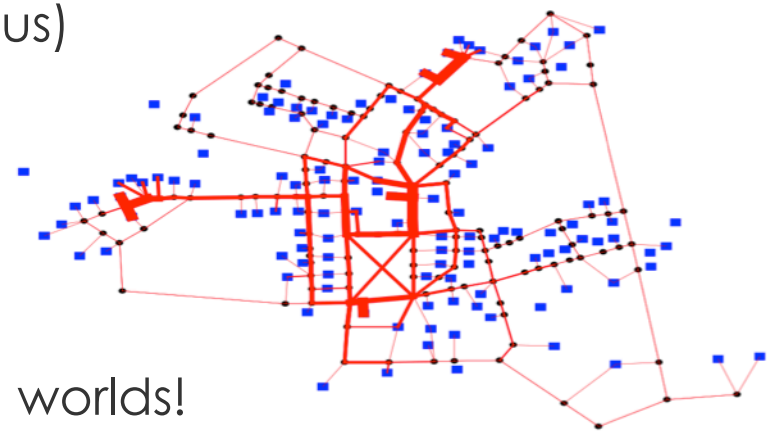
WANT!!
1. Large-scale
2. Accurate
3. Detailed

Adhoc/vehicular Network Design

Infrastructure Deployment

Getting Real Mobility Data is Hard

- Paying users does not scale
- WiFi registration traces (univ campus)
  - Sparse/incomplete
- Cellular base-station pings
  - Sparse sample, privacy issues
- Getting mobility traces from virtual worlds! (mobility traces from SecondLife, WOSN'08)

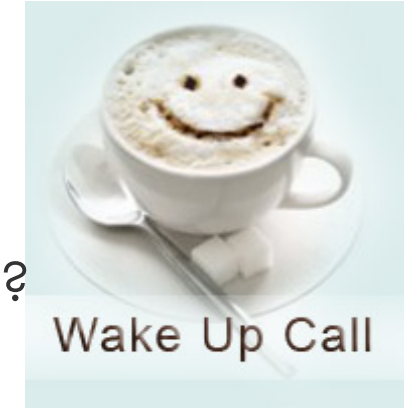


A New Hope: Geo-Social Traces

- "Check-in" to locations, share with friends
- As of Sept. 2013, Foursquare has 40 million users, 4.5 Billion Checkins, 1.3 million businesses

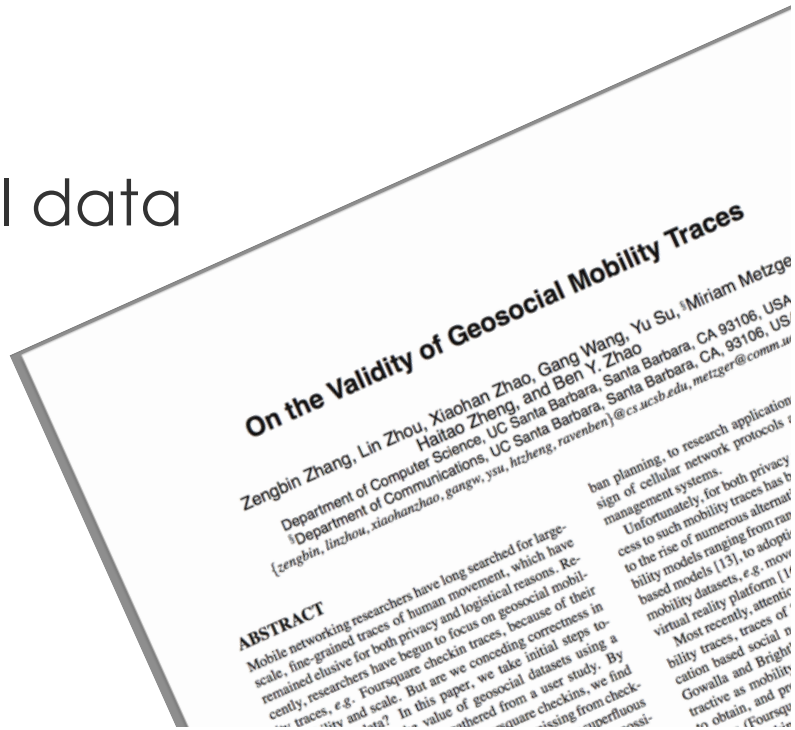
Is This the Answer?

- Geo-social traces increasing available (e.g. FourSquare)
  - Larger scale datasets, crawl-able or public (Twitter)
  - Researchers use data for:
    inferring friendships, predicting human movement, urban planning, CDN design
- But is this data representative?
  - Do incentives to check-in cause bias?
  - Are check-ins a small sample of real mobility?
- Our work: try to understand "reality"
  - How representative are Geosocial Location traces?
  - What is the impact of any errors on applications?
  - Can we try to compensate for these errors?



The Rest of This Talk

- Motivation: Why
- Methodology: How
  - Measuring human mobility
  - Getting user participation
- Data Analysis: What
  - (In-)consistencies of geosocial data
  - Application level impact
- Conclusion


On the Validity of Geosocial Mobility Traces

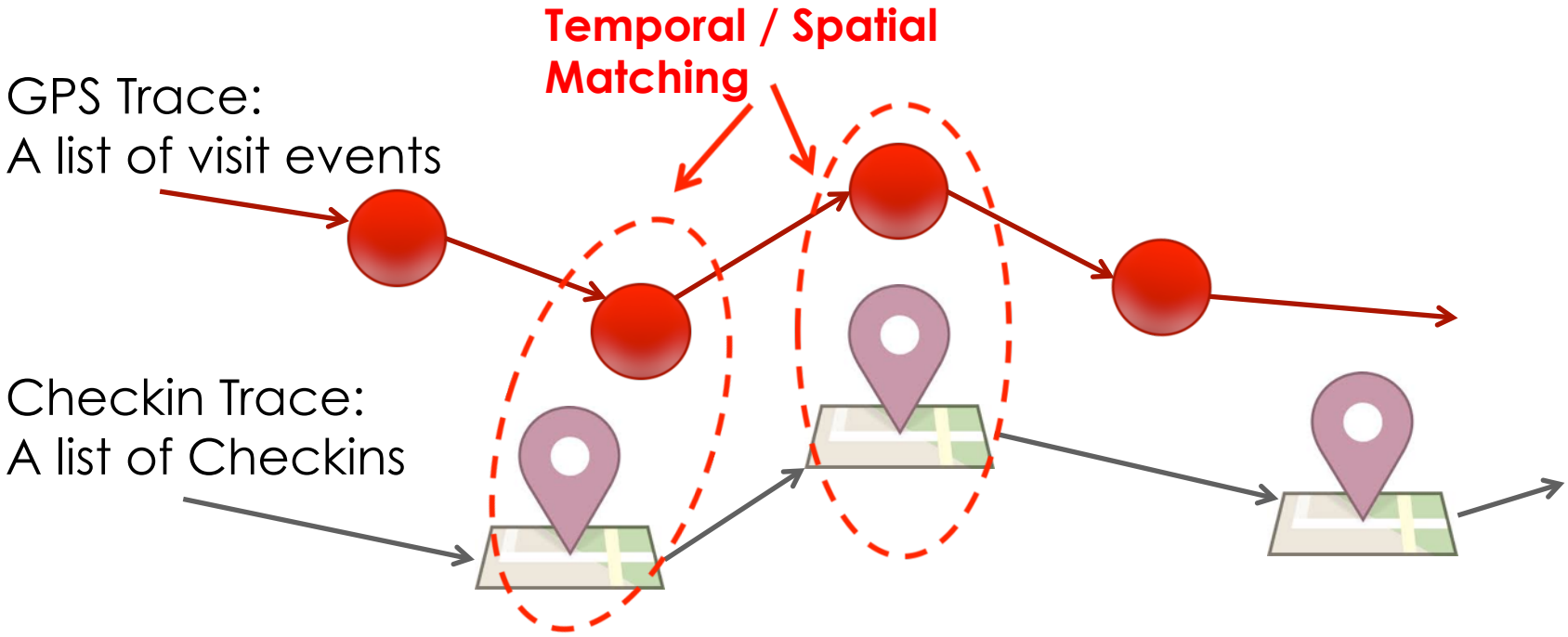
Zengbin Zhang, Lin Zhou, Xiaohan Zhao, Gang Wang, Yu Su, §Miriam Metzge
Haitao Zheng, and Ben Y. Zhao
Department of Computer Science, UC Santa Barbara, Santa Barbara, CA 93106, USA
§Department of Communications, UC Santa Barbara, Santa Barbara, CA, 93106, US
{zengbin, linzhou, xiaohanzhao, gangw, ysu, htzheng, ravenben}@cs.ucsb.edu, metzger@comm.u

ABSTRACT
Mobile networking researchers have long searched for large-scale, fine-grained traces of human movement, which have remained elusive for both privacy and logistical reasons. Recently, researchers have begun to focus on geosocial mobility traces, e.g. Foursquare checkin traces, because of their ...ity and scale. But are we conceding correctness in ...ta? In this paper, we take initial steps to... value of geosocial datasets using a ...thered from a user study. By ...square checkins, we find ...issing from check-... superfluous ...

ban planning, to research application ... sign of cellular network protocols ... management systems.
Unfortunately, for both privacy ... cess to such mobility traces has b... to the rise of numerous alternat... bility models ranging from ran... based models [13], to adopti... mobility datasets, e.g. mov... virtual reality platform [1...
Most recently, attentio... bility traces, traces of ... cation based social n... Gowalla and Bright... tractive as mobility... to obtain, and pr...

The Plan

- Gather simultaneous traces of GPS location and FourSquare check-ins
  - Match check-ins to GPS "events" for consistency



Challenge: Collecting parallel GPS and check-in data from the same set of users

Data Collection

- We need: reasonable user size + fine granularity data
  - Data gathering via Android/iOS smartphone apps
- Functionality
  - Record all FourSquare checkins via API
  - Record GPS once per minute → GPS Trace
    - Turn on WiFi/accelerometer when GPS signal is weak
    - An algorithm to detect *visit* events
    - *Visit* event: user stay @ one place for 6min+
  - Reminder feature to attract 4Sq users
    - All checkins from reminders removed from dataset (only 10% of check-in events)

Data Collection Continued…

- It's not easy
  - Getting user participation
  - Challenges: privacy, power consumption
- First thought: crowdsourcing!
  - Pay to install application, leave the rest to user
  - Many did not provide data
  - For the rest: fake checkins dominate user activity

Surprising result: highly biased data **#fail**

- Fallback plan: organic app adoption
  - Boosted awareness with ads on popular websites
  - Slow adoption rate over 5 months
  - A slow wait for users …

Datasets and Processing

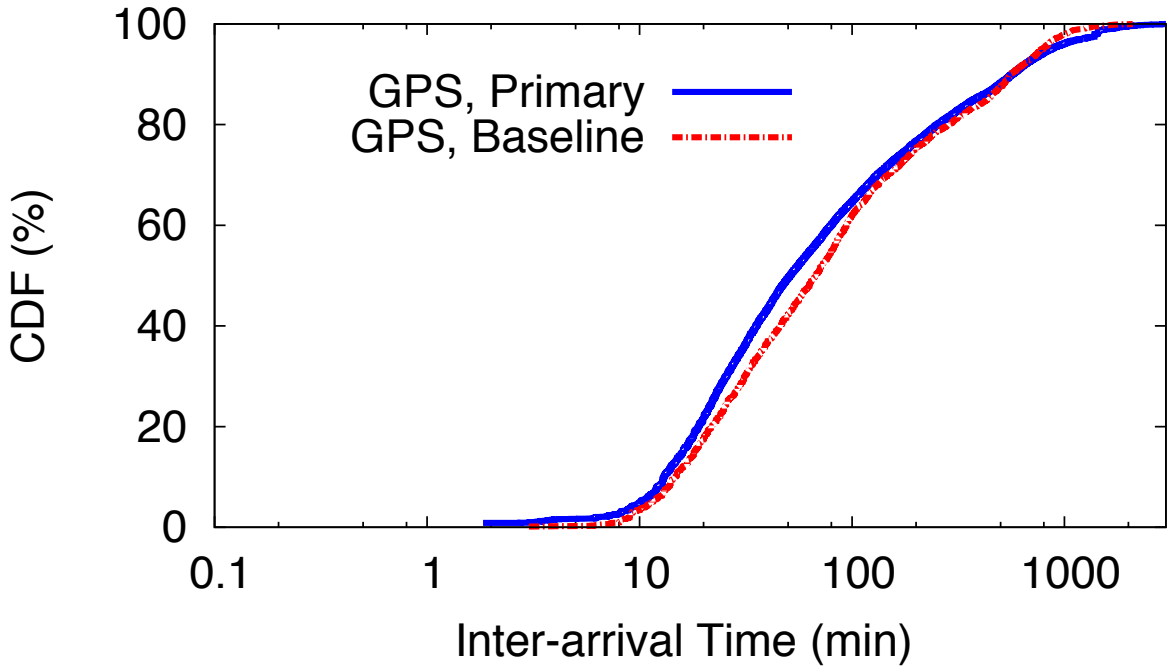
- "*Primary*" dataset
  - Normal Foursquare users who installed our app
- "*Baseline*" dataset
  - Student volunteers from UCSB Communications Dept.
    - Less likely to be affected by Foursquare incentives

	# of users	Avg. days per user	# of checkins	# of GPS visits	GPS Points
Primary	244	14.2	14K	31K	2.6M
Baseline	47	20.8	665	6.3k	558K

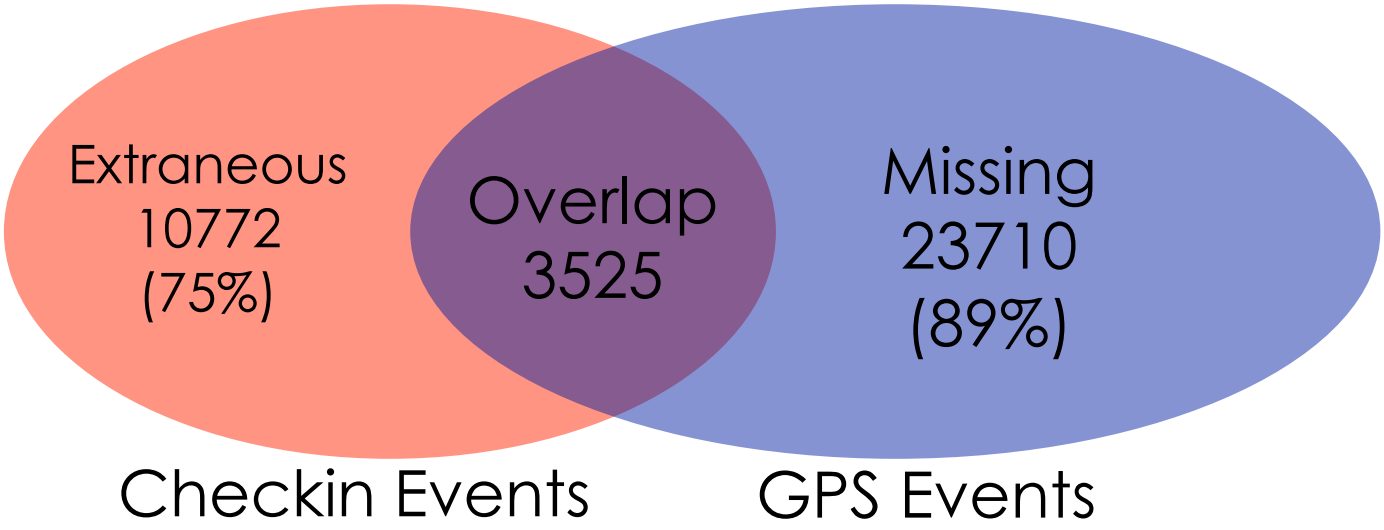
- "Visit Events" from GPS trace: any stationary location for 6+ mins
- Matching events from two traces
  - Conservative: match check-in to closest GPS event, within 500m in distance, within 30 mins before/after check-in

Limited Data Validation

- Correctness of our GPS event generation
  - 92% of Baseline's check-ins match GPS events
- Are app users consistent with UCSB students
  - High level consistency in physical mobility patterns



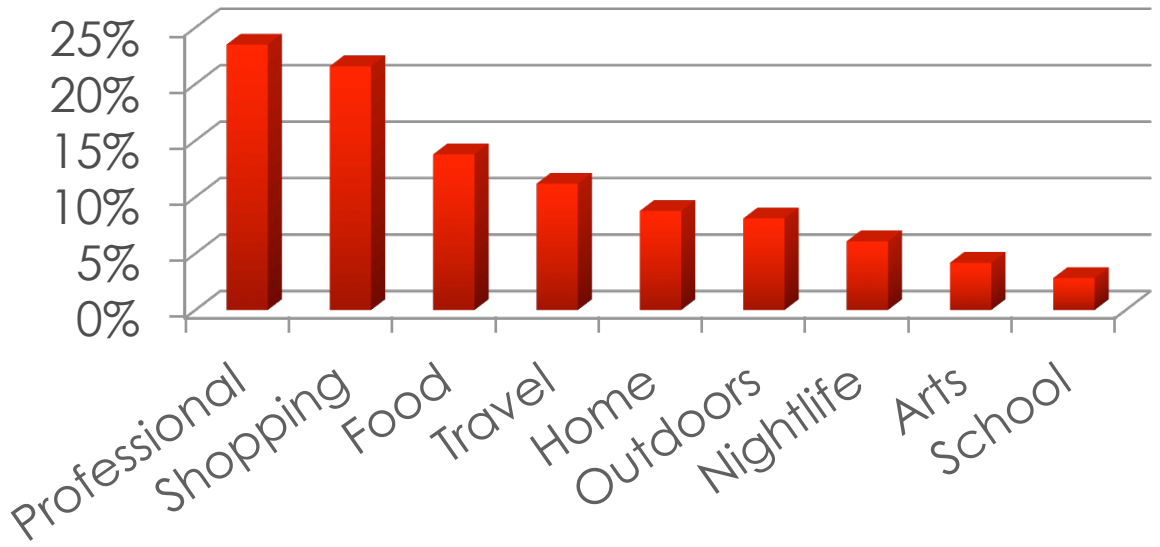
FourSquare vs. Physical Mobility



75% Checkins are extraneous ("cheating")

89% of visited GPS locations have no checkins

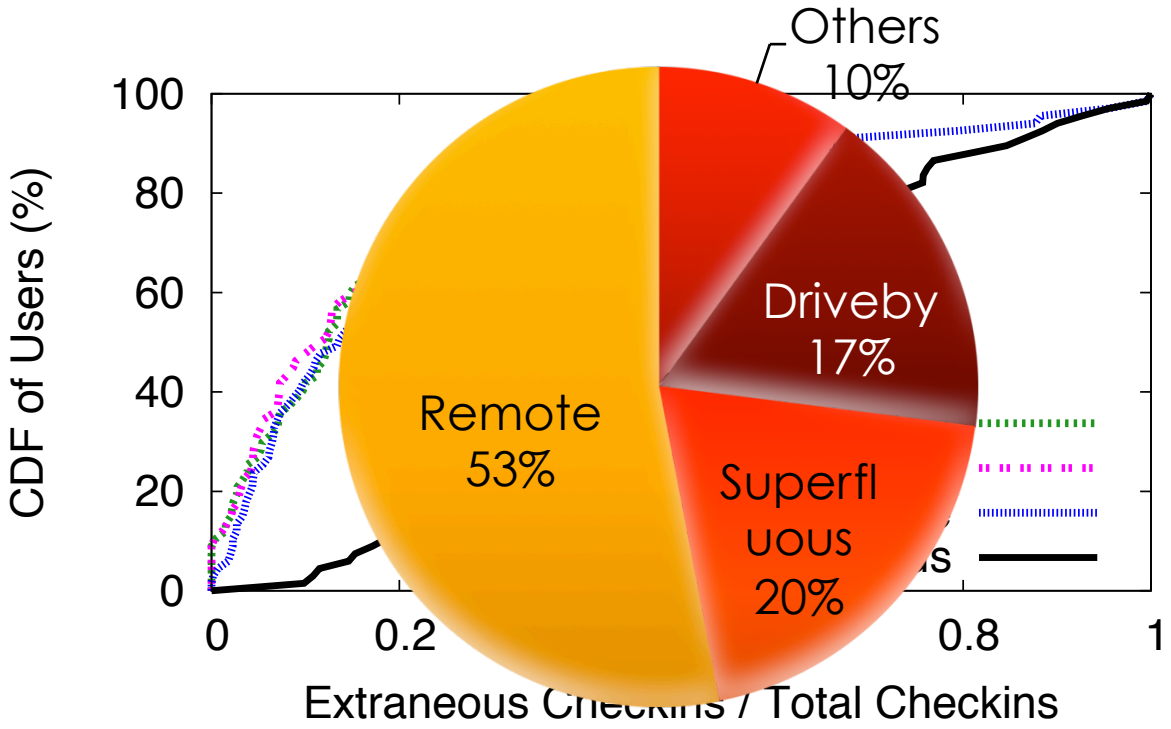
Missing Places / Events



- Routine activities *missing* from FourSquare check-ins
  - Professional (i.e. work), Shopping, Food (eating out)
  - Intuition: users tend to ignore "boring" or potentially private locations

Extraneous Check-ins

- Behavior prevalent across all users
  - Widespread: for 20% of users, 80+% events are extraneous
  - Not black/white: removing users responsible for 80% of extra events would also remove 50% of legitimate data



- Types of extraneous check-ins
  - ***Drive-by***: check-in while moving at high speed
  - ***Superfluous***: checking in nearby locations
  - ***Remote***: check-in location >500m away

Are Incentives Responsible?

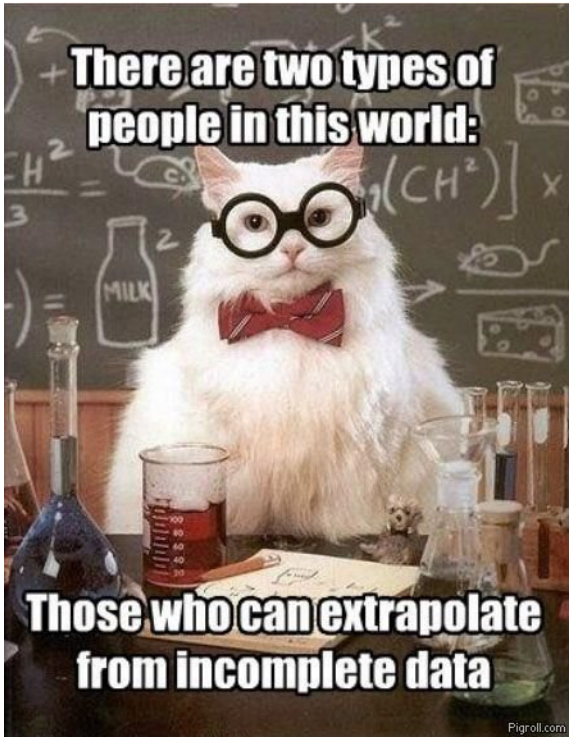
- Study correlation (Pearson's) between per-user rate of extra check-ins and FourSquare user features
  - # of badges, # of mayorships, check-ins/day

- High correlation with Foursquare rewards
  - Rate of remote check-ins *vs.* # of Badges: **0.49**
  - Rate of superfluous check-ins *vs.* # of Mayorships: **0.34**
  - Very intuitive

FourSquare incentives have significant correlation to extraneous check-ins

Key Take-aways

- Large discrepancies between GPS trace and check-ins
  - Large amount of FourSquarecheck-ins are extraneous
  - Majority of real life events missing
- High correlation bet. extra check-ins and FourSquare rewards
- Ongoing work: can we "fix" this?
  - Detecting/removing extraneous checkins
    - Extract useful features
    - Use machine learning to classify
    - Unsupervised methods?
  - Missing check-in extrapolation
    - Temporal and spatial extrapolation



- Thank you!
- Questions?