Session II:

Data and Inference

Jukka-Pekka "JP" Onnela

Associate Professor

Department of Biostatistics

Harvard University

June 5, 2018

OPPORTUNITIES AND CHALLENGES

Opportunities:

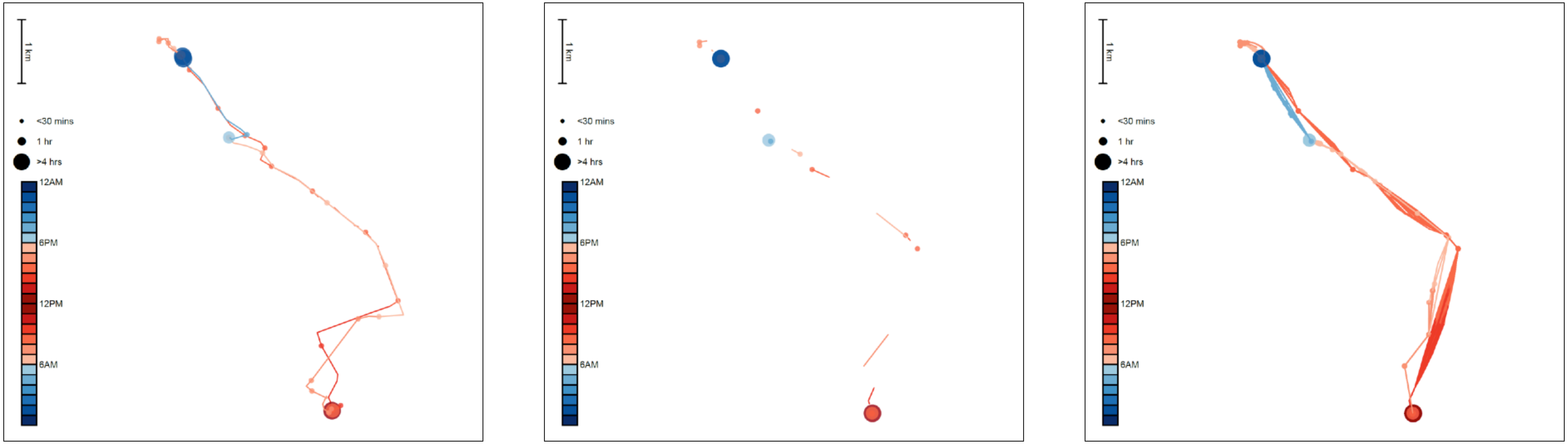
- Scale (N) and length (T) of measurements
- Essentially identical measurements everywhere using smartphones (not so for wearables)
- Building partnerships across industry, academia, government

Challenges:

- Data are very high dimensional and very noisy
- Many analytical and statistical challenges in these early days
- Data standards and reproducibility
- Data security and patient privacy
- Regulatory considerations
- Patient engagement and feedback (possibly useful, can be harmful, constitutes an intervention)
- Integration into clinical care and EHR (information overload)

MOBILITY

- Complete mobility trace vs. simulated missingness
- Typical sampling cycle: on-cycle = 2 mins, off-cycle = 10 mins; 83.3% of mobility trace missing

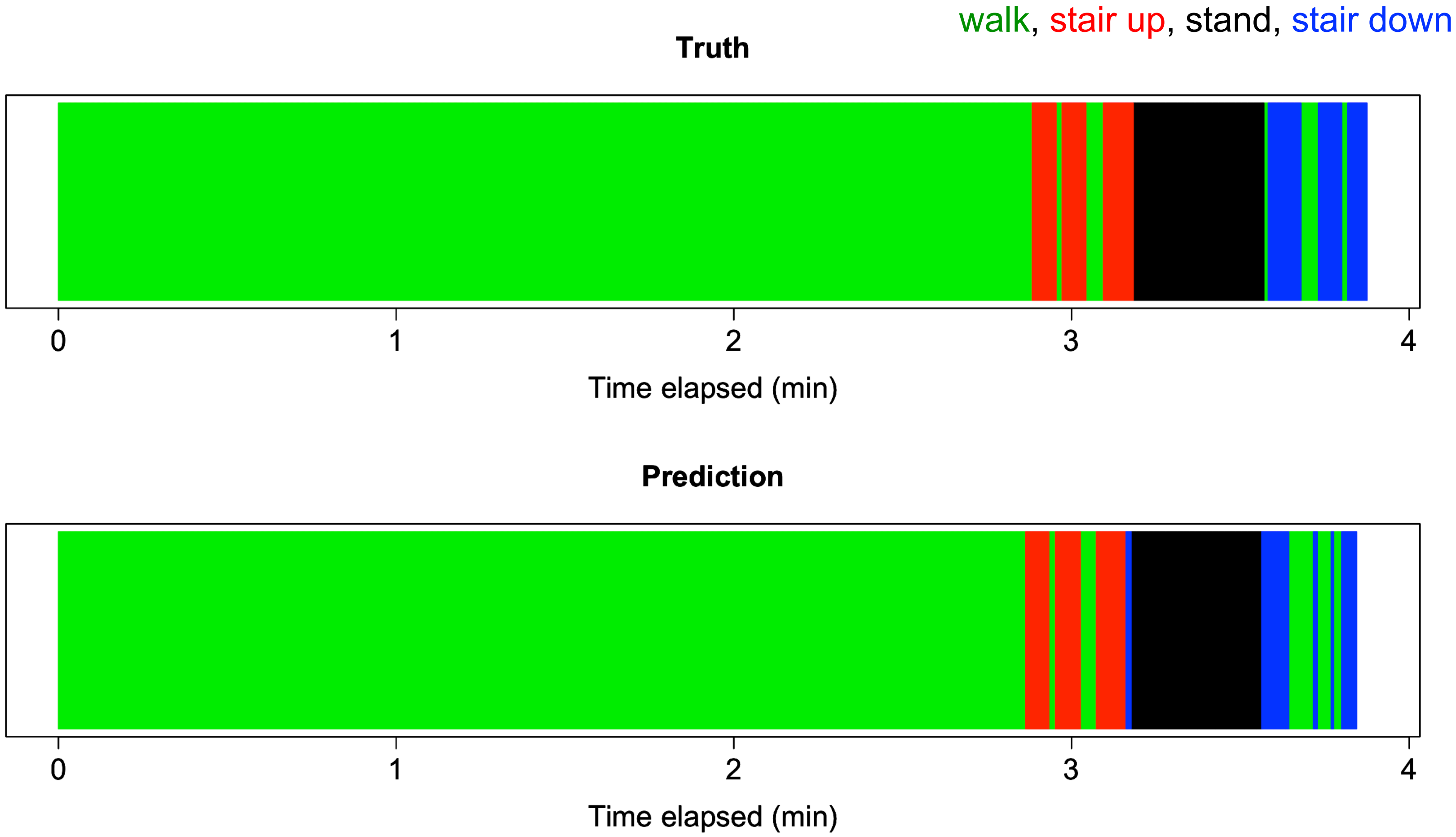


Measures	TL.1	TL.10	TL.20	GL.1	GL.10	GL.20	GLC.1	GLC.10	GLC.20	LI	Truth
StdFlightLen	152.9 ±30.8	125.8 ±10.1	123.2 ±5.5	213.4 ±51.5	205.8 ±36.3	202.7 ±43.5	151.0 ±30.0	134.2 ±8.4	137.1 ±9.0	639.6	223.3
AvgFlightDur	79.0 ±9.3	69.4 ±5.8	68.8 ±5.6	119.0 ±17.9	115.2 ±13.4	113.5 ±13.7	65.4 ±10.5	57.2 ±4.1	60.0 ±5.1	340.6	77.0

Join work with Ian Barnett

PHYSICAL ACTIVITY

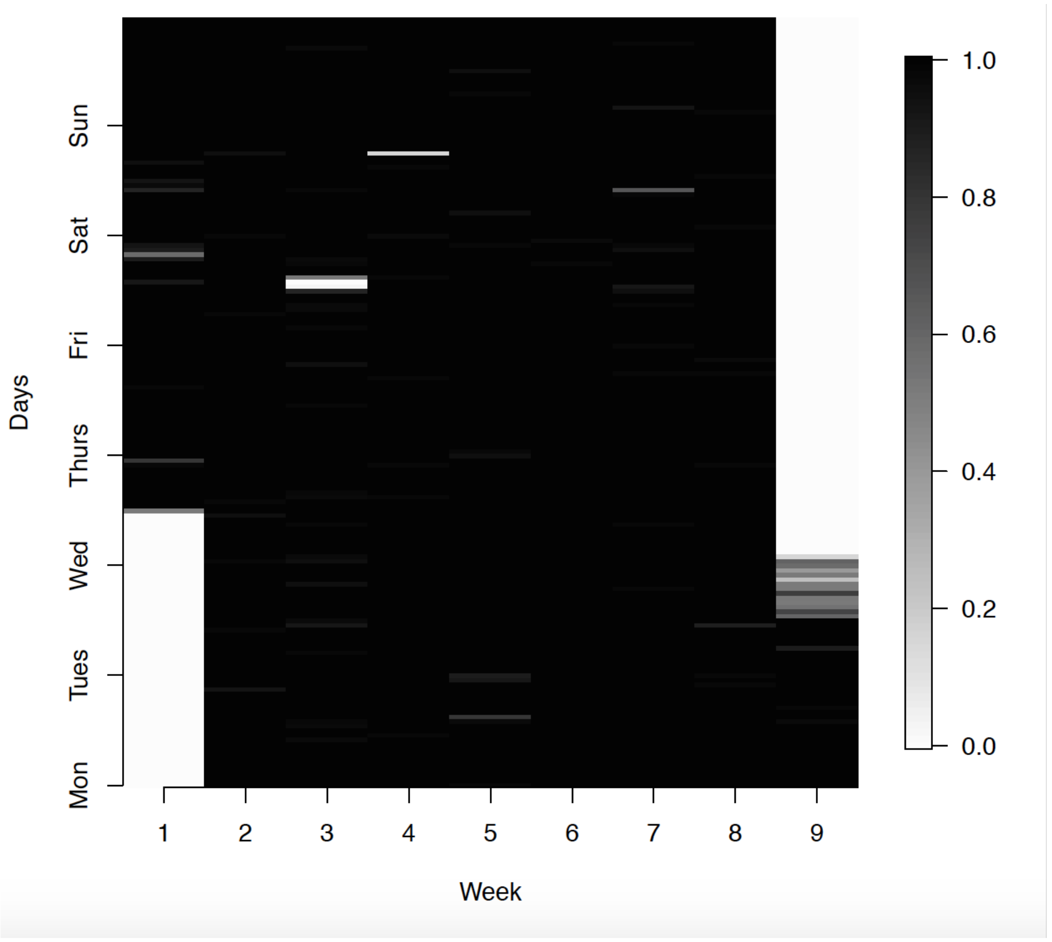
- Activity segmentation using data from gyroscope
- Less data than accelerometer, more even sampling, relevant for activity classification

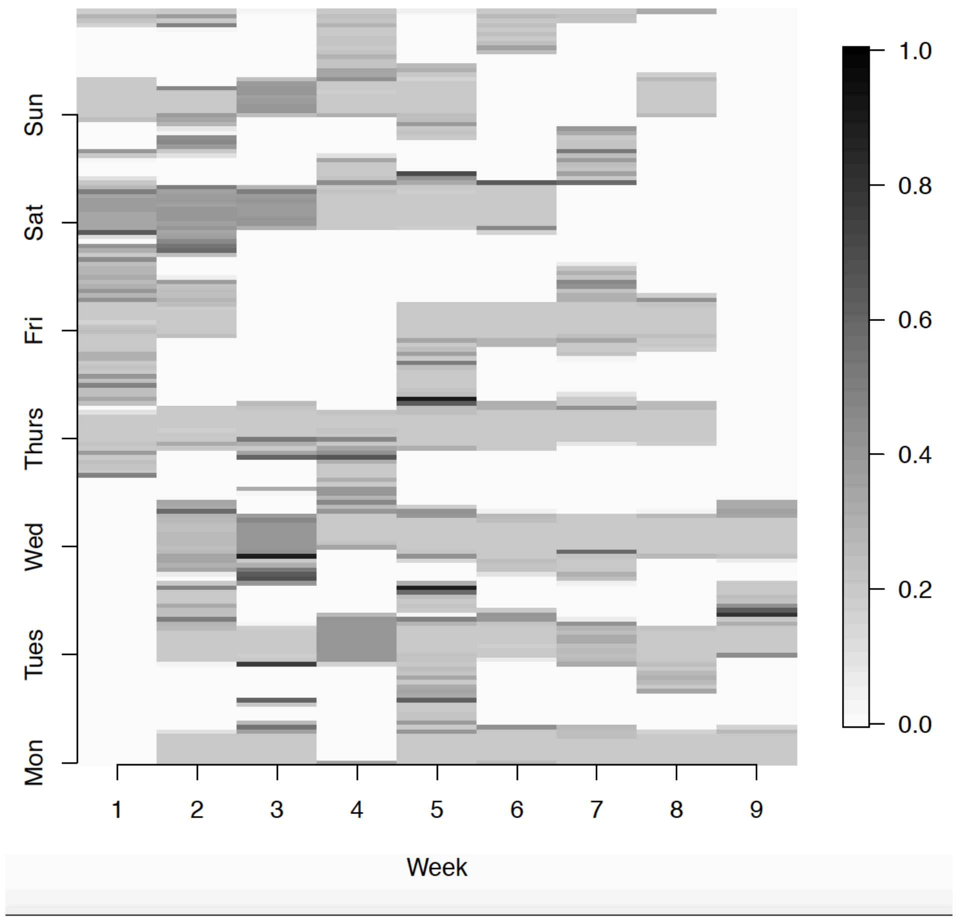


Join work with Emily J Huang

PHYSICAL ACTIVITY

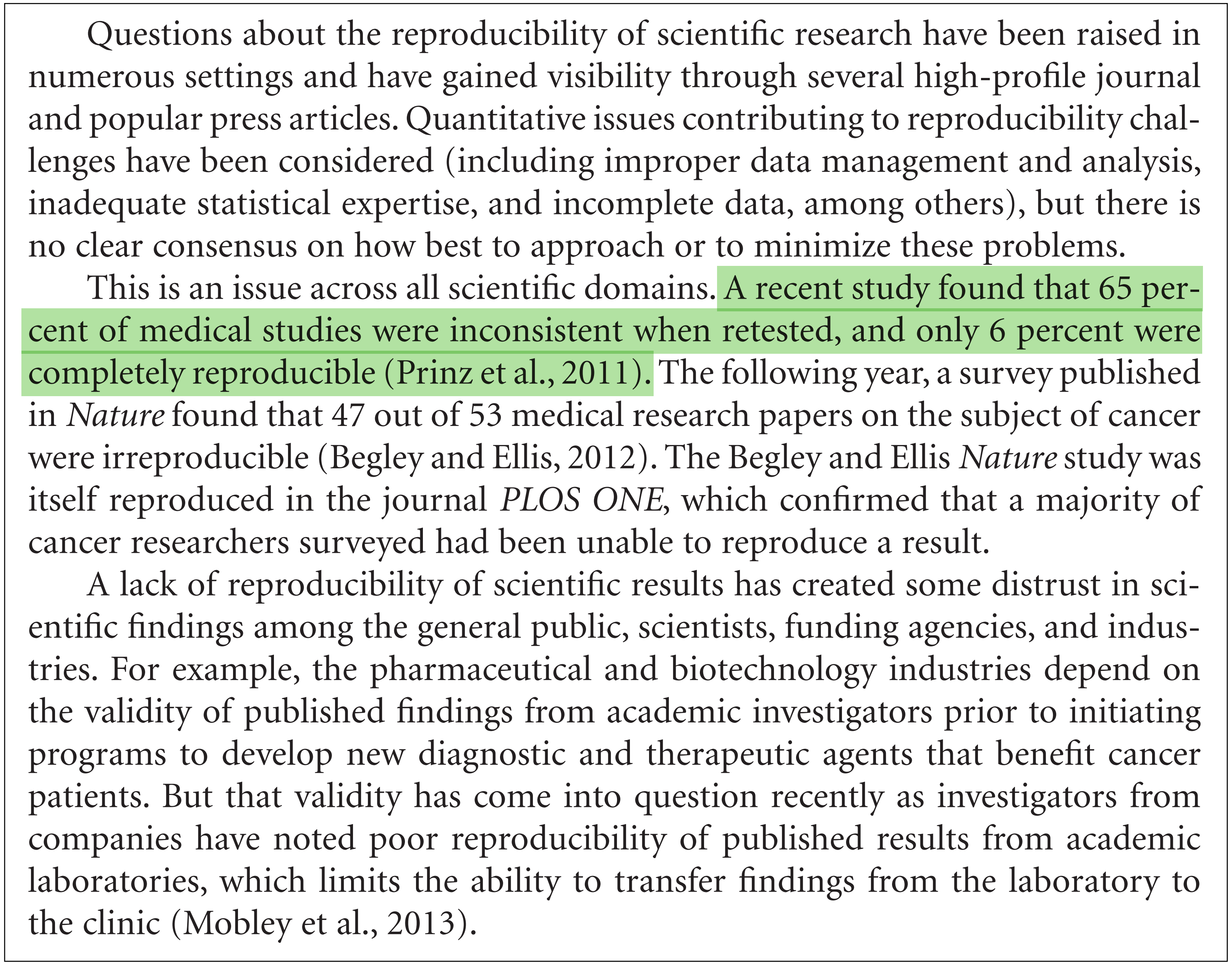
- Accelerometer data has variable sampling rate and coverage
- Propagating uncertainty in data collection to inference is crucial





Join work with Emily J Huang

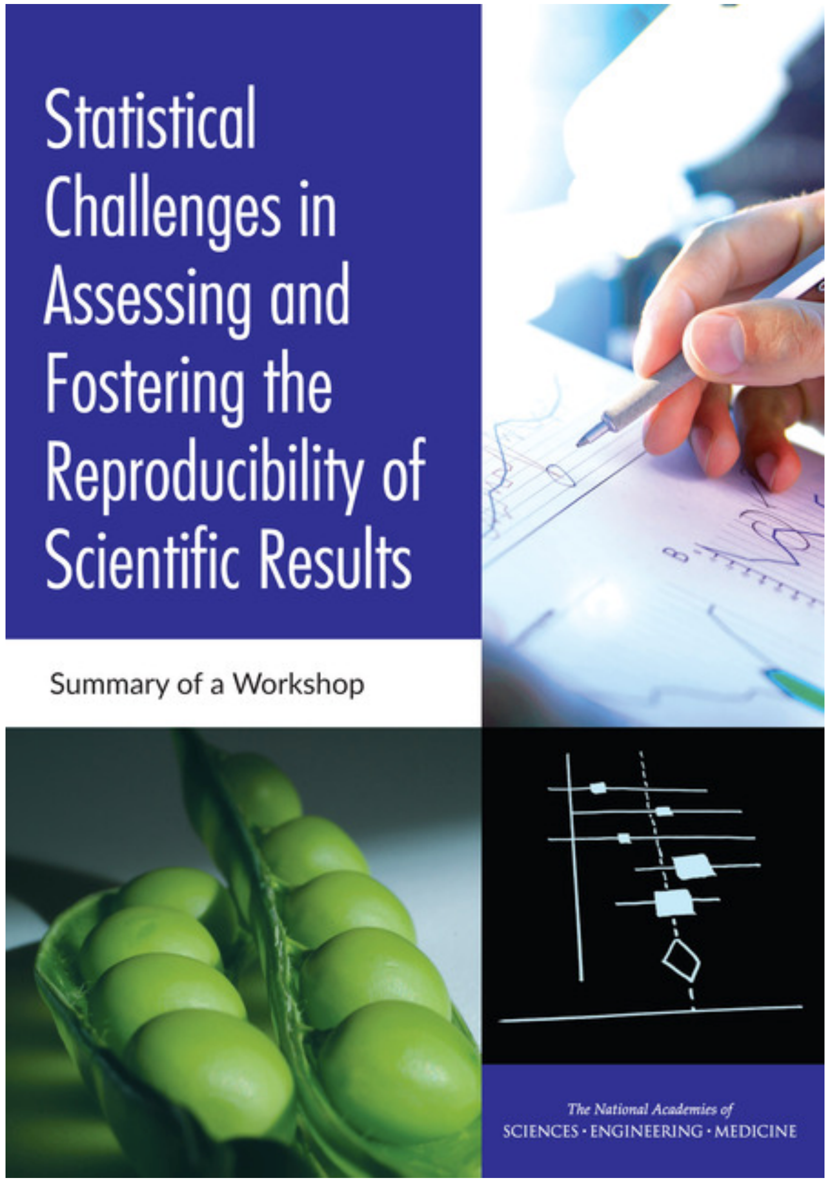
REPRODUCIBILITY



Questions about the reproducibility of scientific research have been raised in numerous settings and have gained visibility through several high-profile journal and popular press articles. Quantitative issues contributing to reproducibility challenges have been considered (including improper data management and analysis, inadequate statistical expertise, and incomplete data, among others), but there is no clear consensus on how best to approach or to minimize these problems.

This is an issue across all scientific domains. A recent study found that 65 percent of medical studies were inconsistent when retested, and only 6 percent were completely reproducible (Prinz et al., 2011). The following year, a survey published in *Nature* found that 47 out of 53 medical research papers on the subject of cancer were irreproducible (Begley and Ellis, 2012). The Begley and Ellis *Nature* study was itself reproduced in the journal *PLOS ONE*, which confirmed that a majority of cancer researchers surveyed had been unable to reproduce a result.

A lack of reproducibility of scientific results has created some distrust in scientific findings among the general public, scientists, funding agencies, and industries. For example, the pharmaceutical and biotechnology industries depend on the validity of published findings from academic investigators prior to initiating programs to develop new diagnostic and therapeutic agents that benefit cancer patients. But that validity has come into question recently as investigators from companies have noted poor reproducibility of published results from academic laboratories, which limits the ability to transfer findings from the laboratory to the clinic (Mobley et al., 2013).



Statistical Challenges in Assessing and Fostering the Reproducibility of Scientific Results

Summary of a Workshop

The National Academies of
SCIENCES · ENGINEERING · MEDICINE

Statistical Challenges in Assessing and Fostering the Reproducibility of Scientific Results (Free PDF available online)

OPPORTUNITIES AND CHALLENGES

Opportunities:

- Scale (N) and length (T) of measurements
- Essentially identical measurements everywhere using smartphones (not so for wearables)
- Building partnerships across industry, academia, government

Challenges:

- Data are very high dimensional and very noisy
- Many analytical and statistical challenges in these early days
- Data standards and reproducibility
- Data security and patient privacy
- Regulatory considerations
- Patient engagement and feedback (possibly useful, can be harmful, constitutes an intervention)
- Integration into clinical care and EHR (information overload)