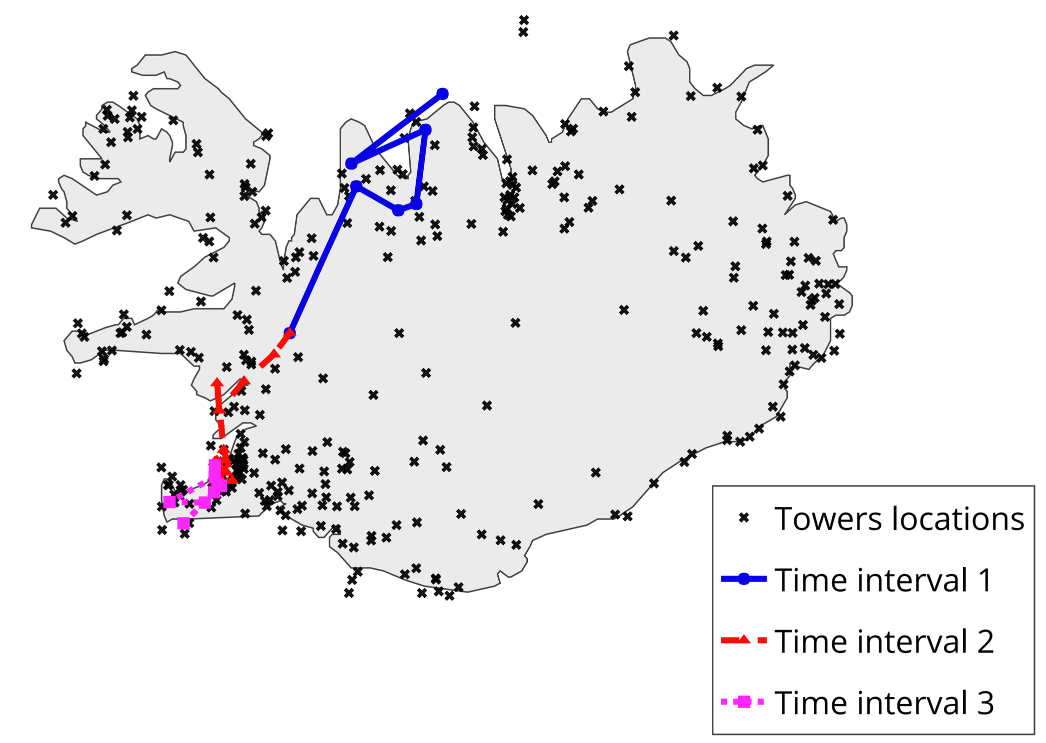
Tracking Behavioral Alterations via Mobile Phone Data

Derek Onken[1], Thorgeir Karlsson[2], Atli Einarsson[2], Congzeng Song[1], Leon Danon[3], Ymir Vigfusson[1]
[1]Math & Computer Science, Emory University [2] Reykjavik University [3] University of Exeter

The Data

- A large mobile network operator supplied their billing data for October 2008 to 2012. We focus on the 2009-2010 records during the H1N1 outbreak.
- The Centre for Health Security and Communicable Disease Control (CHS-CDC) in Iceland provided the date of diagnosis for a patient who displayed symptoms of influenza.
- Data Protection Authority (Personuvernd) approved the anonymizing process.

Cellphone towers and movement inference



subject	object	time	In	call	tarif	tariftype	units	towerid	lat	lon
98937	52674	2010-09-17 10:34:46	t	f		PREP	0	719	65.679166	-18.092559
4197	89504	2010-05-06 16:07:24	t	t	GIN	PREP	7	287	66.152133	-18.903783
51993	607	2010-09-29 01:47:50	f	t	GGSM7	POST	25	617	65.66145	-18.10765

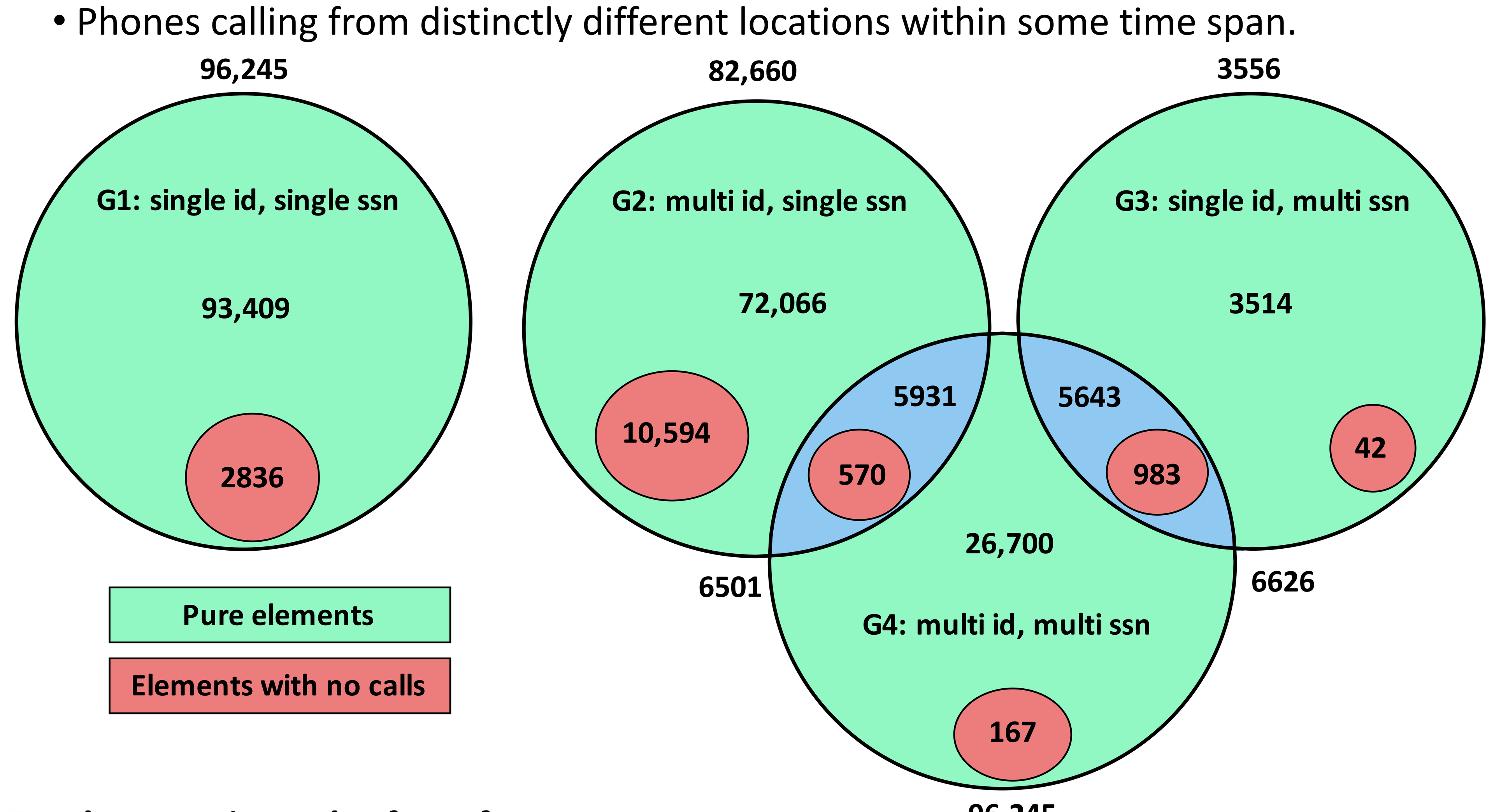
Calls table. Each line represents information related to the phone id labeled subject.

ssn_no	famely_no	in_nat_reg	cust_type	first_record	illness
41486	41486	1	Person	2009-10-04	Influenza
2732	24003	1	Person	2009-10-28	Influenza
749	40780	1	Person	2009-05-25	Influenza

Health table. Each line represents one diagnosis.

Quantifying the Groups

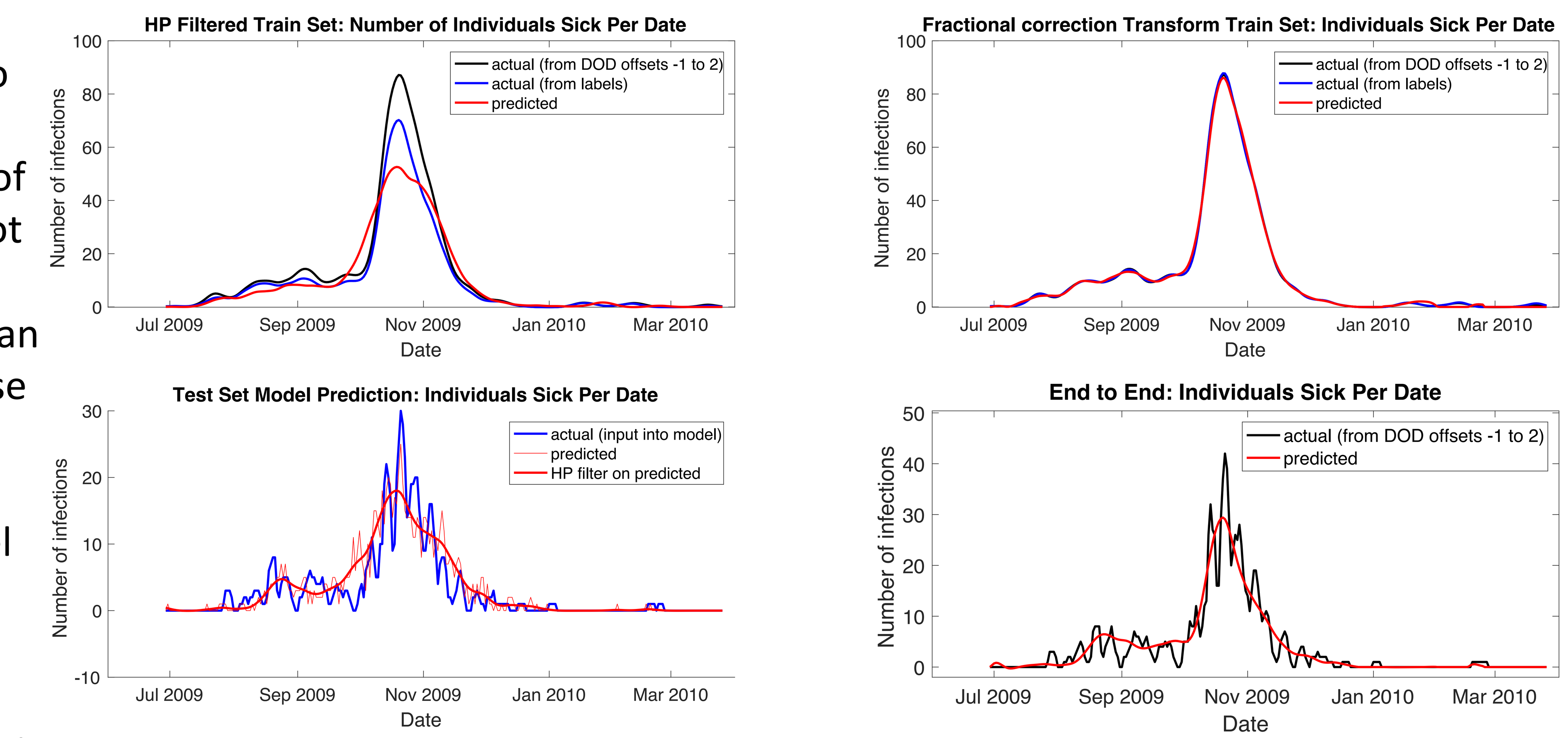
- Not every phone or ssn represents one individual (families, people with a company phone, etc.)
- Data split into four groups based on (id,ssn) pairs. Groups 3 and 4 are primarily companies.
- Group 2: How many people exist here?
- Determine distinct people based on:
  - Sequential use of a phone (disjoint sequential use implies separate individuals),
  - Phones calling each other within a ssn (assume those are 2 separate people),
  - Phones calling from distinctly different locations within some time span.



Total counts in each of our four groups.

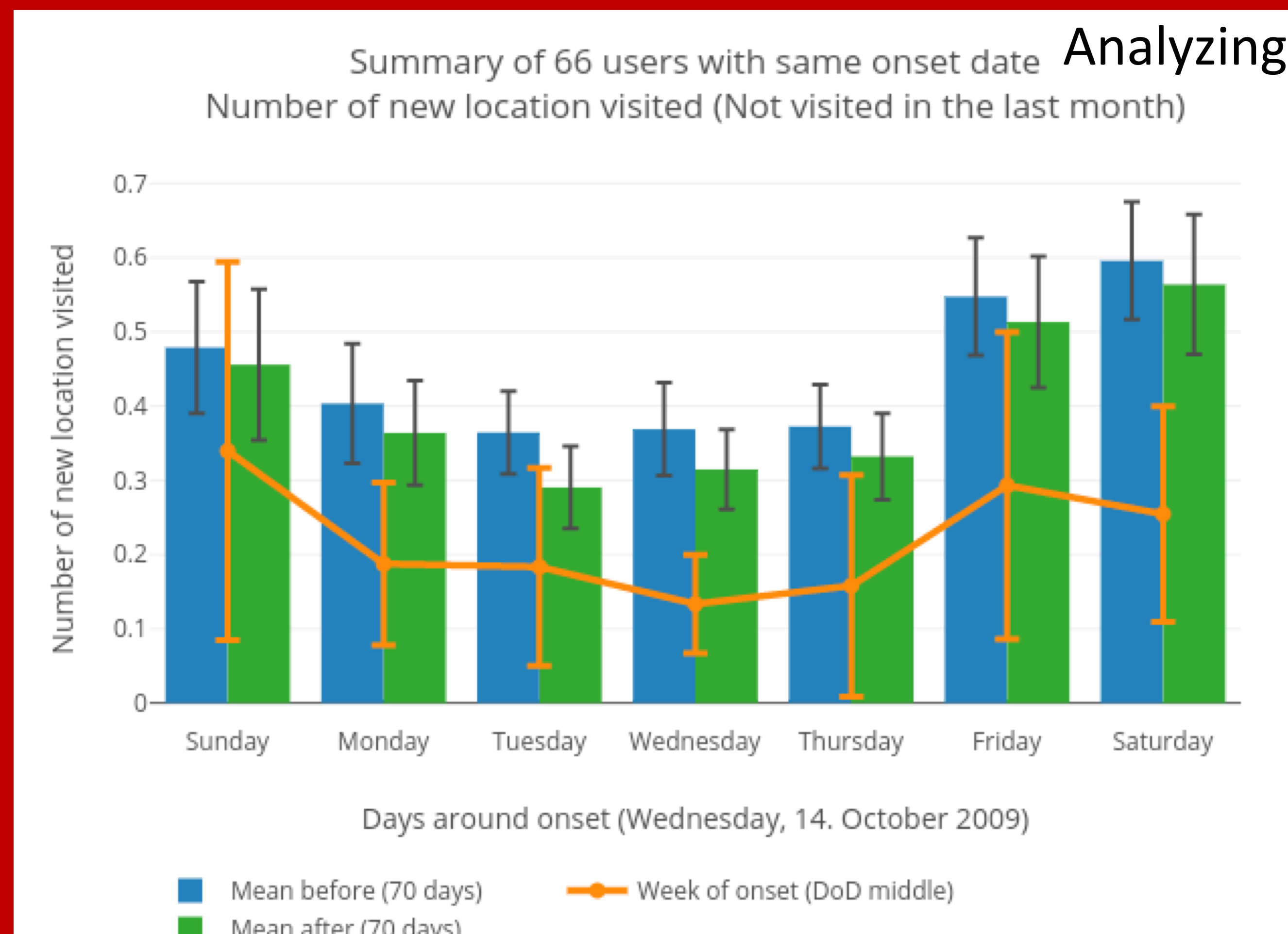
Baseline Model

- For our baseline model, we applied naïve linear regression to the single id, single ssn group.
- We pass week-long sequences of features with a label of sick or not into our regression.
- Our input (blue curve) is less than the full data (black curve) because some individuals lack density of phone-use data.
- We use the output of the model on the training set to define a correctional transform.
- We apply the transform to the output of the model on testing set, smoothing it with Hodrick-Prescott.

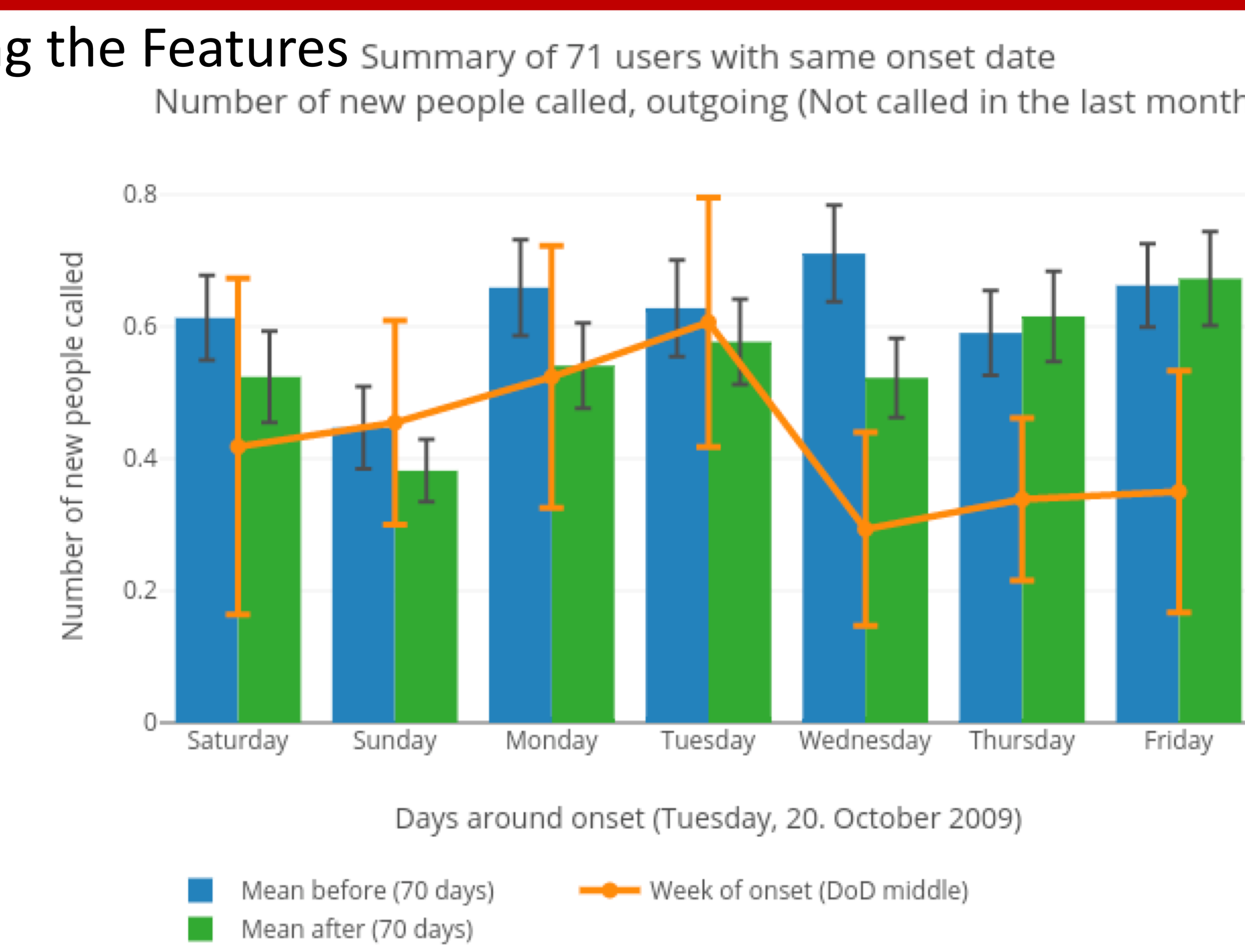


Plotting our model's output of the epidemic curve vs. our "ground truth"

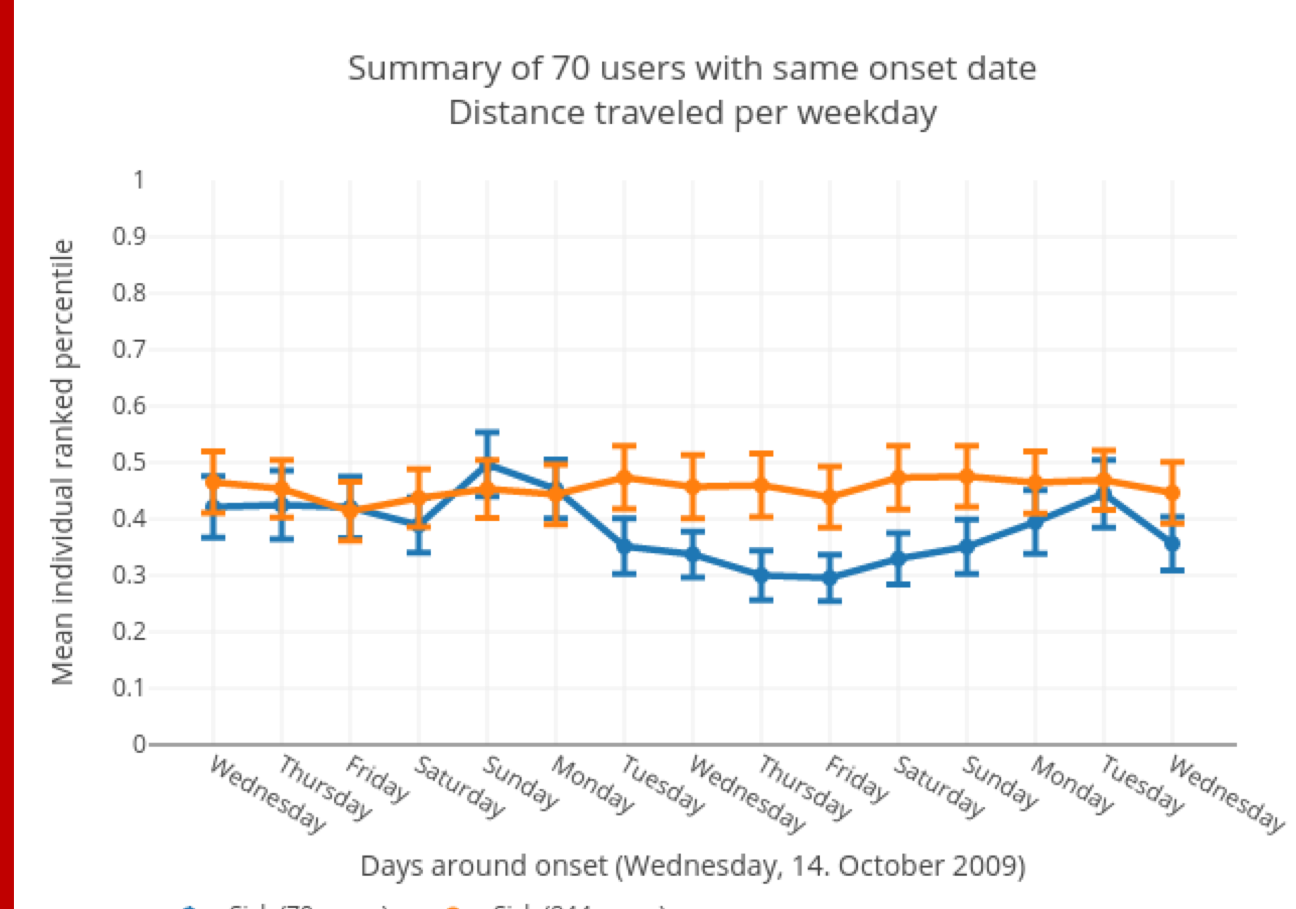
Analyzing the Features



Sick people visit fewer new locations than when healthy



Sick individuals call new/weaker contacts less



Sick individuals move around less than when healthy

- Differences in behavior occur between weekdays and weekends.
- To account for this, we compare sick individuals to other sick individuals diagnosed on the same date.
- Compare mean of feature in the ten weeks before, the ten weeks after, and the week of illness.
- Do this for every feature. Some plots of the features with most clear distinction are displayed here.
- This is only on the training and validation sets of the single id-single ssn group.