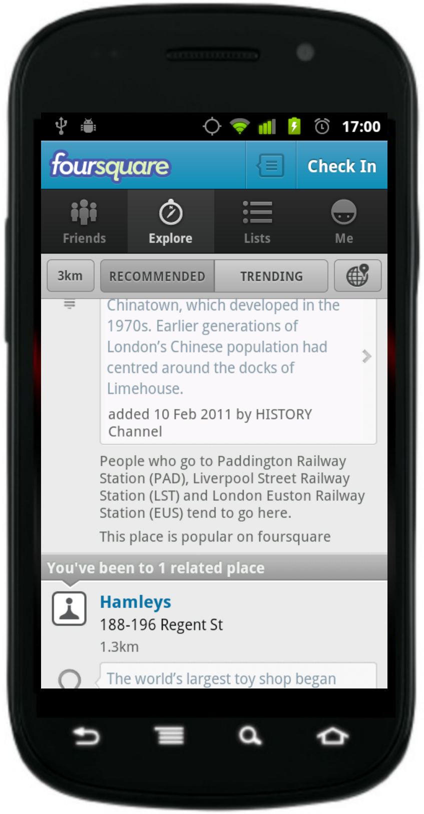
Putting Ubiquitous Devices' Data to Use

@neal_lathia
University of Cambridge
January 23, 2013


21:32
Tube Star
LONDON
Log In:
With Twitter
Anonymously


oyster
Transport for London
Issued subject to conditions – see over



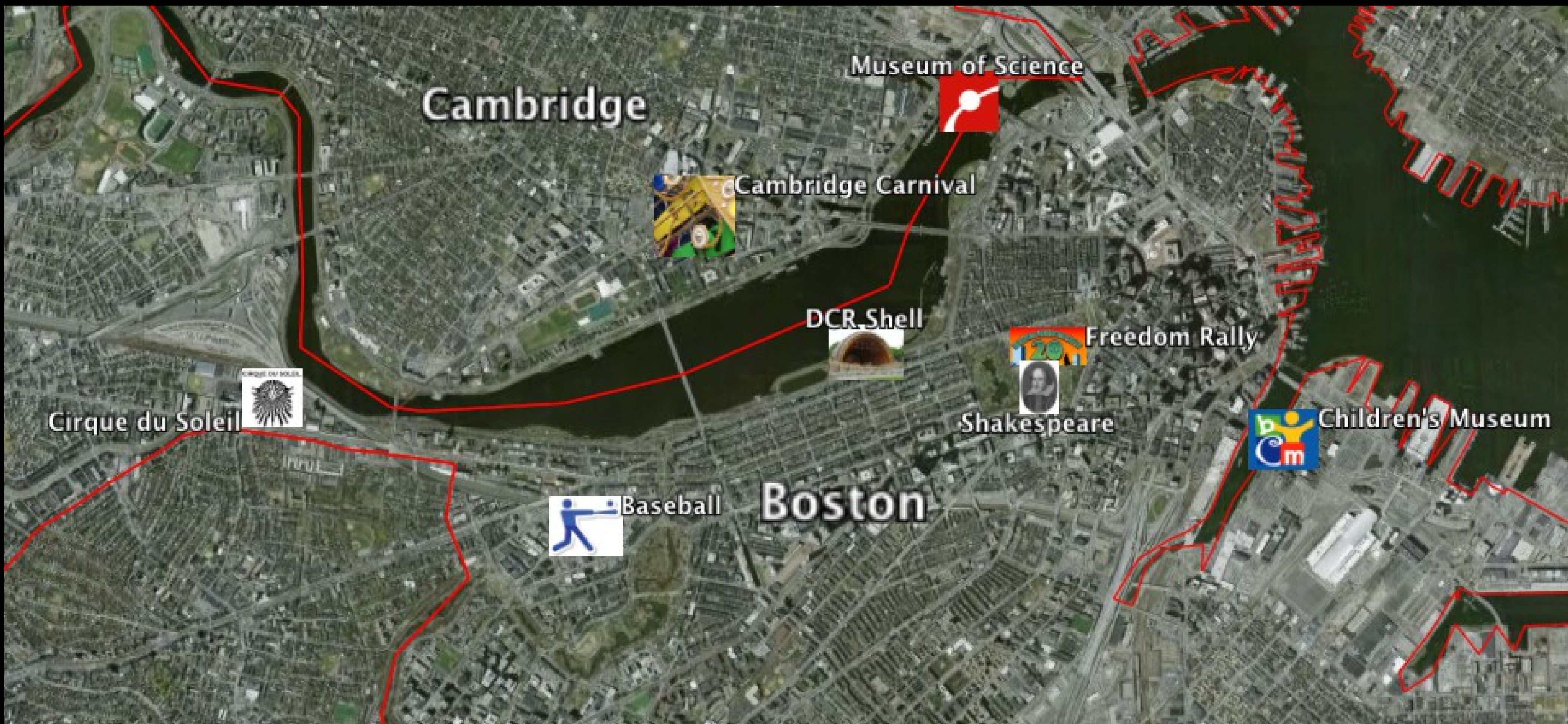
what data can we collect?

recommender systems
aim to match users to items that will be of interest to them

recommender systems

aim to match ~~users~~ mobility profiles to ~~items~~ social events that will be of interest to them use mobility data to recommend social events

(1) infer attendance at events
(2) recommend (test 6 different algorithms)
(3) evaluate recommendation quality


Museum of Science
Cambridge
Cambridge Carnival
DCR Shell
Freedom Rally
Shakespeare
Children's Museum
Cirque du Soleil
Baseball
Boston

task

(1) get users' data, split temporally
(2) run algorithm that outputs recommendations...
(3) evaluate the quality of the recommendations

algorithms

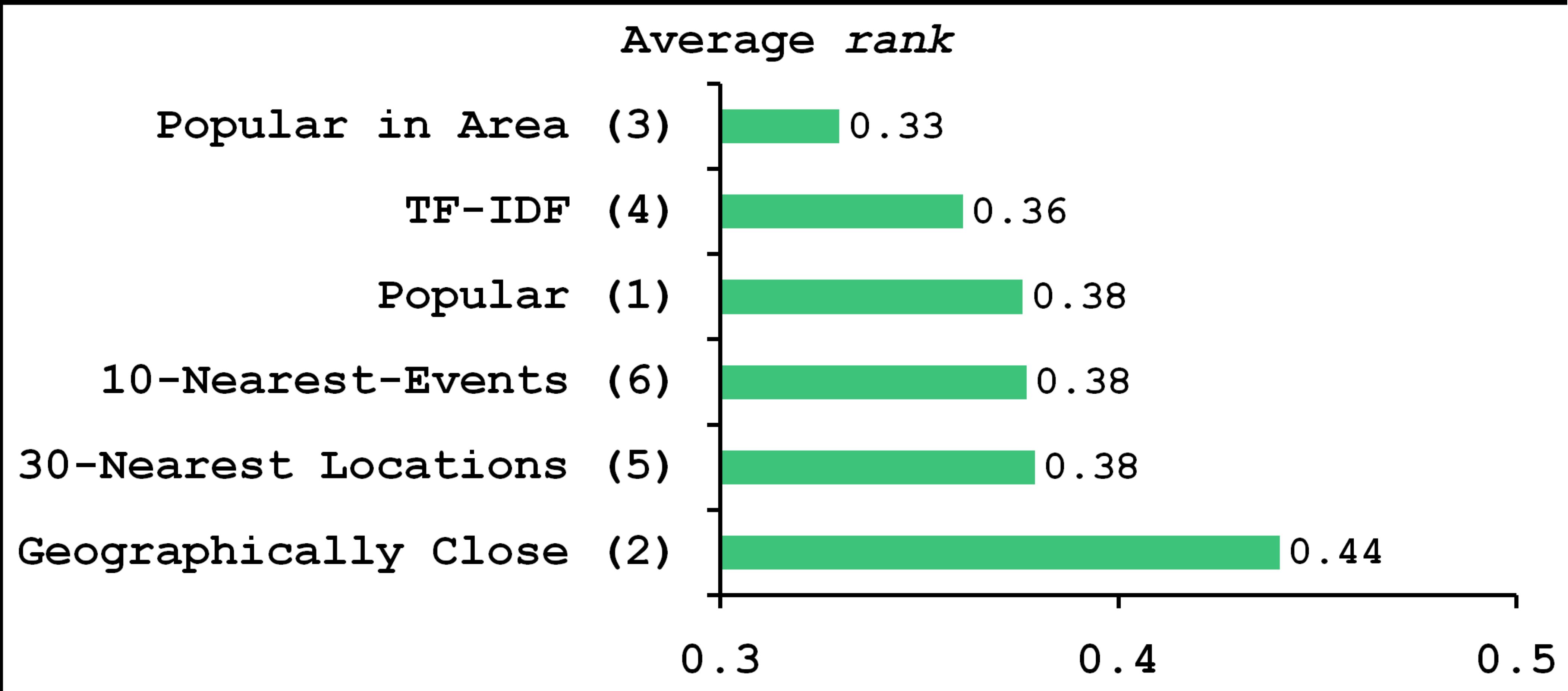
(1) popular events (in the city)
(2) geographically close
(3) popular events (where you live)
(4) TF-IDF
(5) k-Nearest Locations
(6) k-Nearest Events

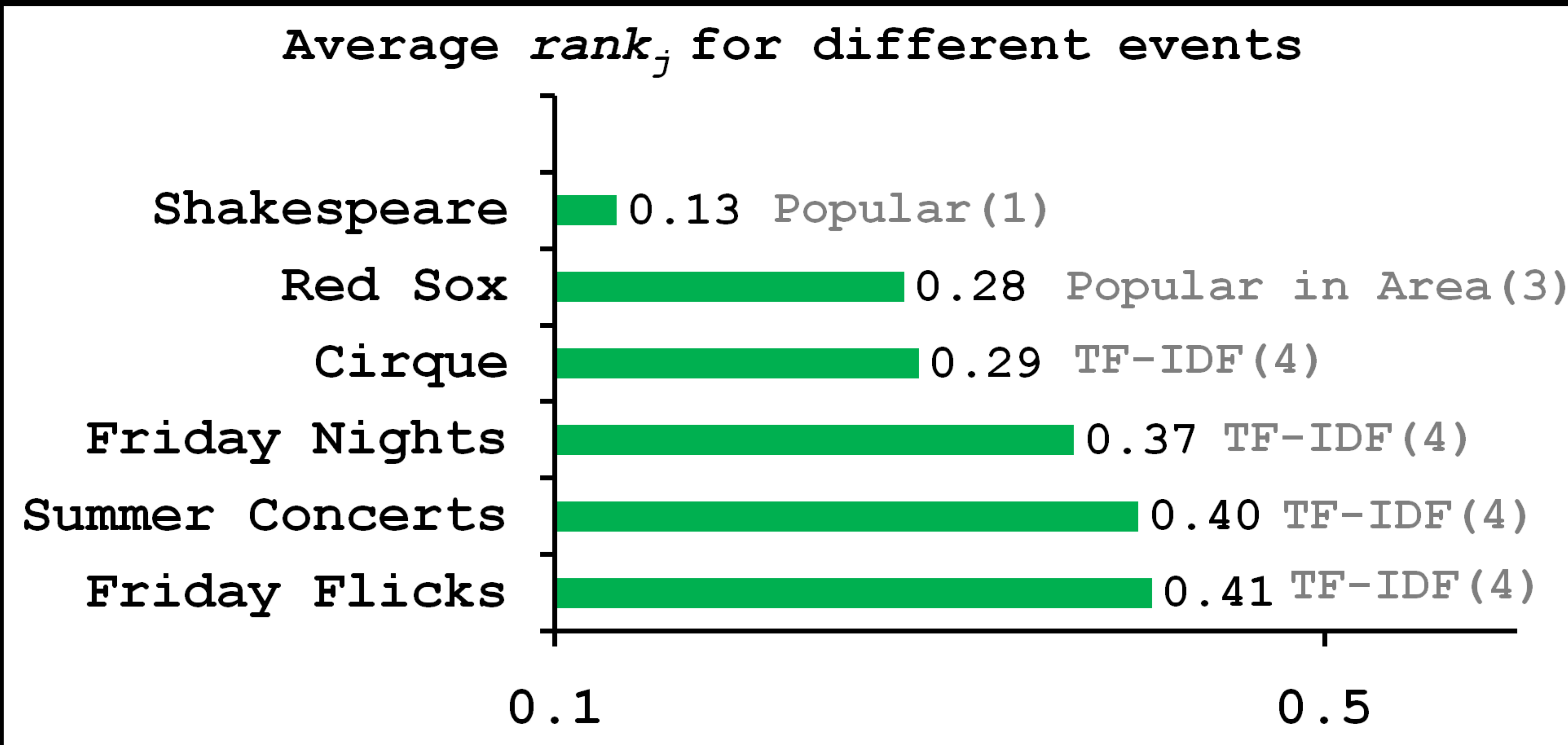
what is a good recommendation?

what is a good recommendation?

evaluate by **ranking:** are the events you went to 'near' the top of the recommendation list?

metric: percentile ranking. small value = good. high value = bad.


Average rank
Popular in Area (3)
0.33
TF-IDF (4)
0.36
Popular (1)
0.38
10-Nearest-Events (6)
0.38
30-Nearest Locations (5)
0.38
Geographically Close (2)
0.44
0.3
0.4
0.5



Average $rank_j$ for different events
Shakespeare
Red Sox
Cirque
Friday Nights
Summer Concerts
Friday Flicks
0.13 Popular(1)
0.28 Popular in Area(3)
0.29 TF-IDF(4)
0.37 TF-IDF(4)
0.40 TF-IDF(4)
0.41 TF-IDF(4)
0.1
0.5

future

how would you use other smartphone sensors to improve recommendations?


oyster
Transport for London
Issued subject to conditions – see over

what tools could we design to help travellers?

sensing mobility:
5%-sample, 2 x 83-days

time-stamped location (entry, exit), modality
payments (top-ups, travel cards)
card-types (e.g., student)

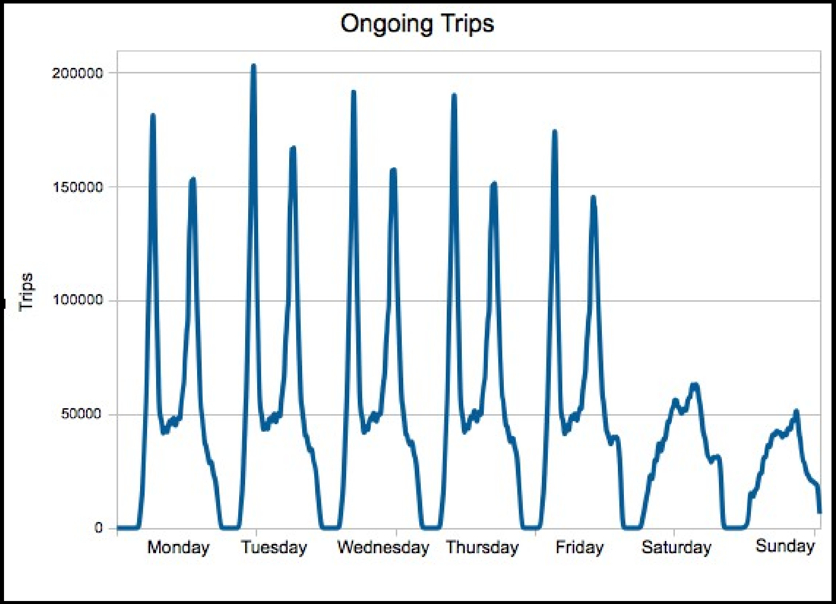
Adult | 18+ student | 16 - 18 | 11 - 15 | 5 -10 | New Deal | Bus & Tram | Railcard | Groups

Adult

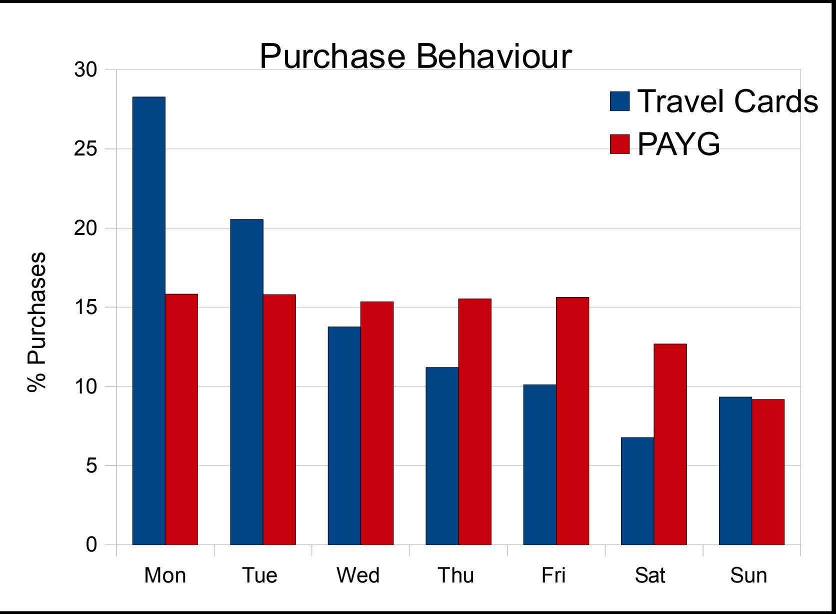
Zone	Cash	Oyster pay as you go				Travelcards				
		Peak single	Off-peak single	Peak price cap	Off-peak price cap	Day Anytime	Day Off-peak	7 Day	Monthly	Annual
Zone 1 only	£4.00	£1.90	£1.90	£8.00	£6.60	£8.00	£6.60	£27.60	£106.00	£1,104
Zones 1-2	£4.00	£2.50	£1.90	£8.00	£6.60	£8.00	£6.60	£27.60	£106.00	£1,104
Euston - Zone 2*	£4.00	£2.00	£1.90	£8.00	£6.60	£8.00	£6.60	£27.60	£106.00	£1,104
Zones 1-3	£4.00	£2.90	£2.50	£10.00	£7.30	£10.00	£7.30	£32.20	£123.70	£1,288
Euston - Zone 3*	£4.00	£2.70	£2.50	£10.00	£7.30	£10.00	£7.30	£32.20	£123.70	£1,288
Zones 1-4	£5.00	£3.40	£2.50	£10.00	£7.30	£10.00	£7.30	£39.40	£151.30	£1,576
Euston - Zone 4*	£5.00	£3.10	£2.50	£10.00	£7.30	£10.00	£7.30	£39.40	£151.30	£1,576
Zones 1-5	£5.00	£4.10	£2.70	£15.00	£8.00	£15.00	£8.00	£47.00	£180.50	£1,880
Euston -	£5.00	£3.80	£2.70	£15.00	£8.00	£15.00	£8.00	£47.00	£180.50	£1,880

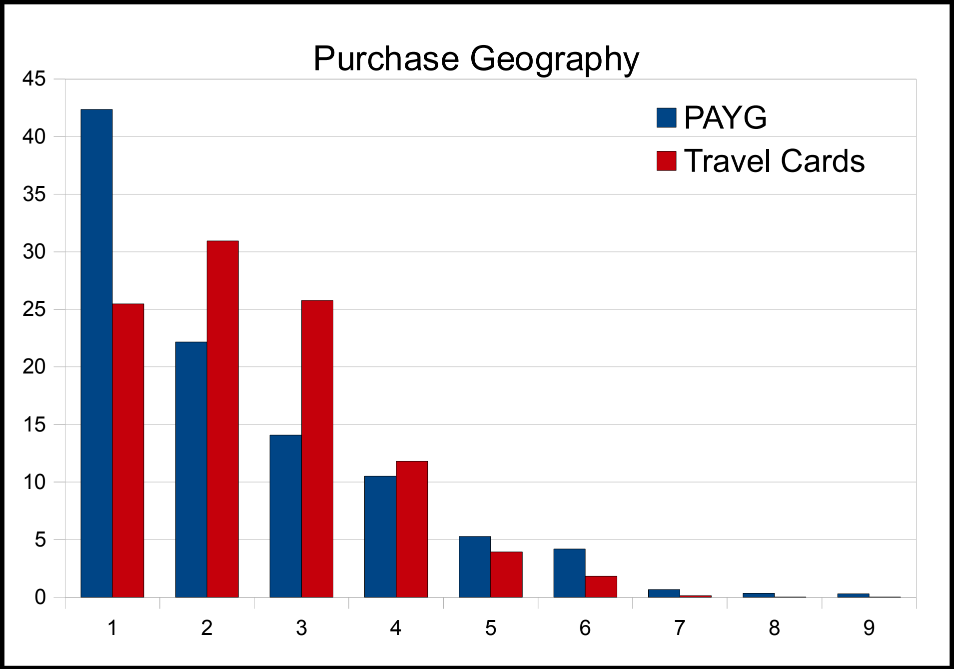
questions

(1) what is the relation between how we **travel** & how we **spend**?
(2) do travellers make the correct **decisions**? (no)
(3) can we help them with **recommendations**? (yes)

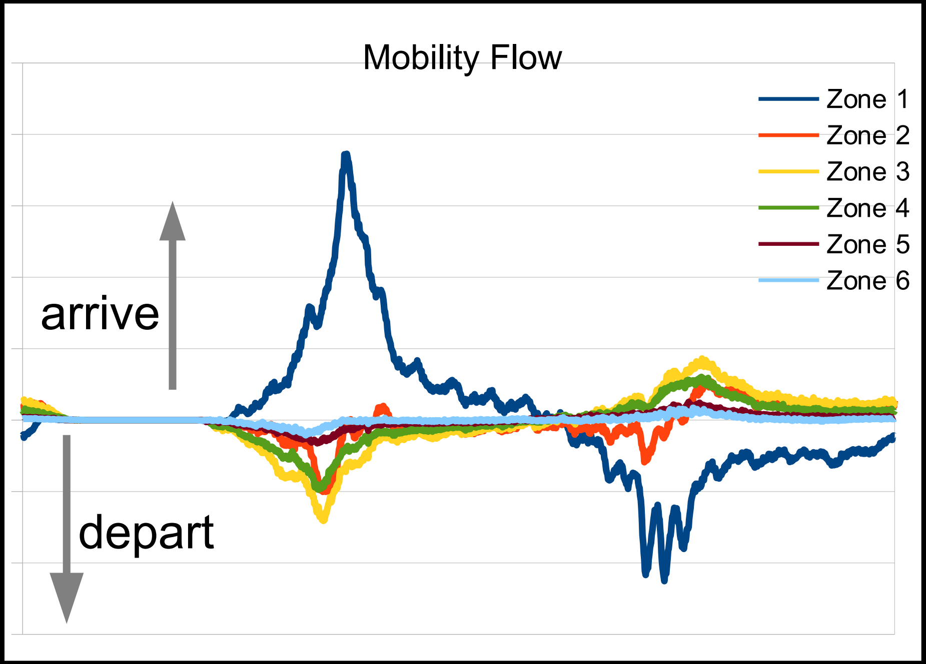


(%)	pay as you go purchases
49.8	< 5 GBP
24.2	5 – 10 GBP
15.5	10 – 20 GBP
(%)	travel card purchases
70.8	7-day travel card
15.8	1-month travel card
11.6	7-day bus/tram pass





Purchase Geography
PAYG
Travel Cards
45
40
35
30
25
20
15
10
5
0
1
2
3
4
5
6
7
8
9



Mobility Flow
Zone 1
Zone 2
Zone 3
Zone 4
Zone 5
Zone 6
arrive
depart

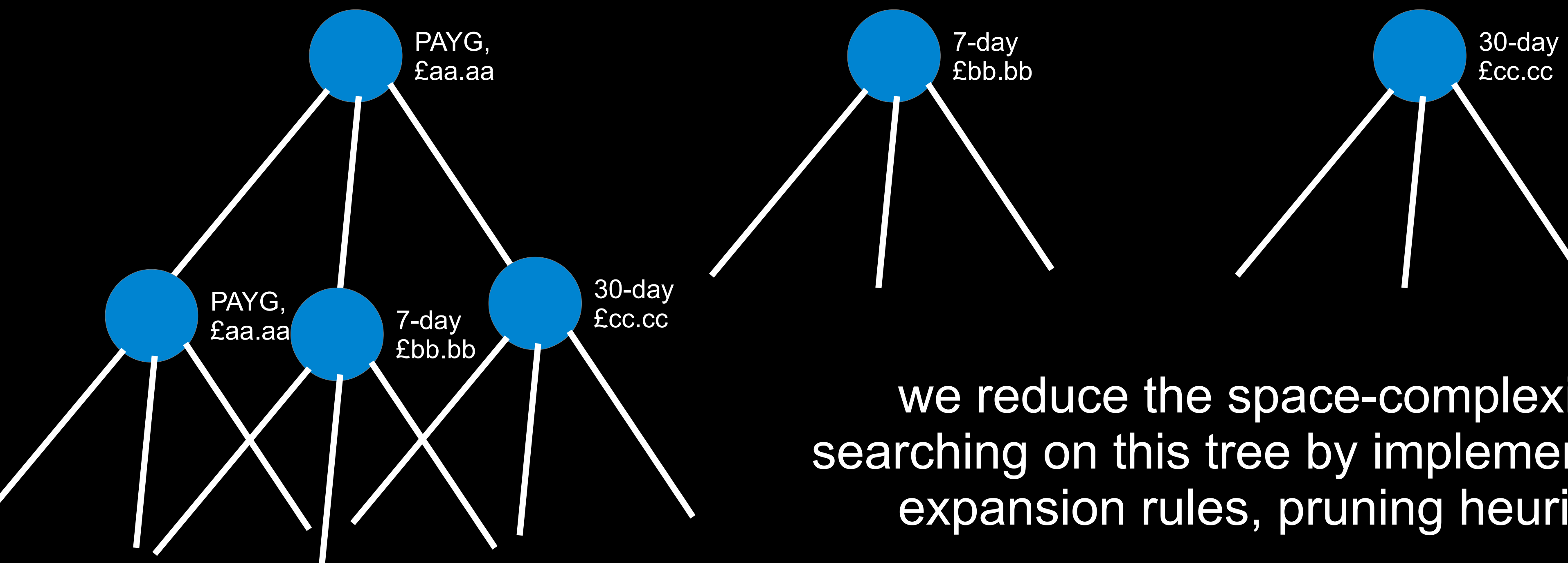
the data shows that:
(a) there is a high regularity in travel & purchase behaviour
(b) travellers buy in small increments and short-terms
(c) most purchases happen upon refused entry

(2) do travellers make the correct decisions?
compare actual purchases to the optimal (per traveller)

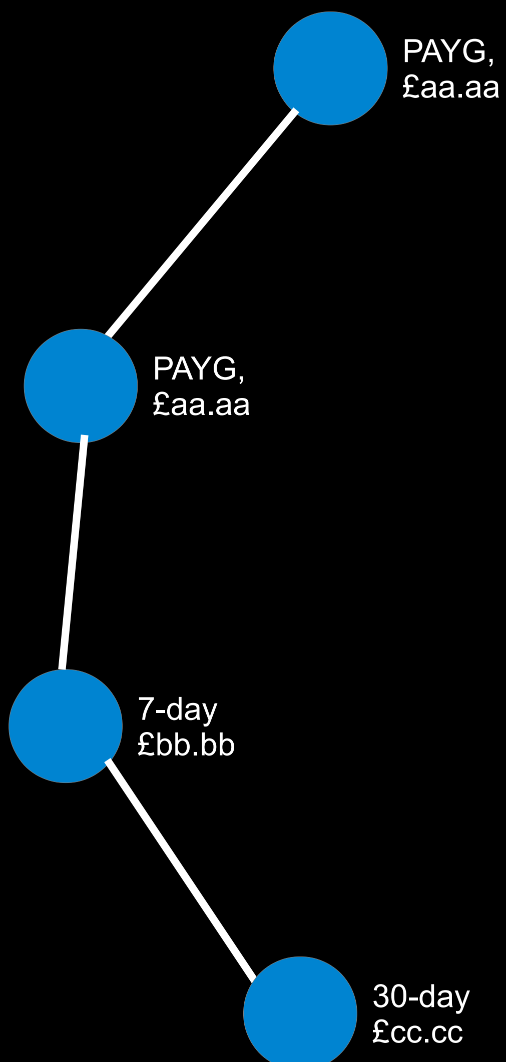
how:
(a) clean data
(b) build & search on a tree ~ sequence of choices

how: build a tree with each user's mobility data
where a node is a **purchase (expire, cost)**
that is expanded when it has expired
(reduced) example:

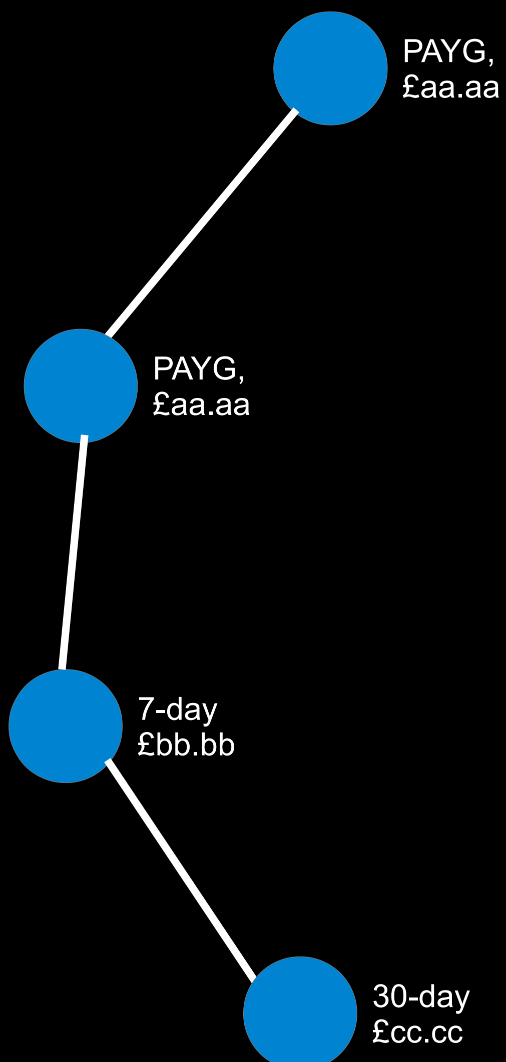
how: build a tree with each user's mobility data
where a node is a **purchase (expire, cost)**
that is expanded when it has expired
(reduced) example:



we reduce the space-complexity of searching on this tree by implementing expansion rules, pruning heuristics the cheapest sequence of fares can then be compared to what the user actually spent



the cheapest sequence of fares can then be compared to what the user actually spent



in each 83-day dataset, the 5% sample of users where overspending by **~ £2.5 million**

An estimate of how much everybody (100%) is overspending during an entire year (365 days) **is thus £200 million**

overspending comes from

(a) failing to predict one's own mobility needs

...but we have observed that mobility is predictable

(b) failing to match mobility with fares (in a complex fare system)

...which is an easy problem for a computer

can we help travellers?

recommender systems
aim to match users to items that will be of interest to them

recommender systems
aim to match ~~users~~ mobility profiles to ~~items~~ fares that will be ~~of interest~~ the cheapest for them

three steps

1. for a given set of travel histories, compute the cheapest fare (by tree expansion)
2. reduce each travel history into a set of generic features, describing the mobility (next slide)
3. train classifiers to predict the cheapest fare given the set of features

we have a set of {d, f, b, r, pt, ot, N} = F

where

d = number of trips
f = average trips per day
b / r = proportion of trips on the bus / rail
pt / ot = proportion of peak & off-peak trips
N = zone O-D matrix
F = cheapest fare (label)

two baselines, three algorithms:

0. baseline – everyone on pay as you go
1. naïve bayes – estimating probabilities
2. k-nearest neighbours – looking at similar profiles
3. decision trees (C4.5) – recursively partitions data to infer rules
4. oracle – perfect knowledge

	Accuracy (%)		Savings (GBP)	
	Dataset 1	Dataset 2	Dataset 1	Dataset 2
Baseline	74.99	76.91	326,447.95	306,145.85
Naïve Bayes	77.46	80.71	393,585.81	369,232.24
k-NN (5)	96.74	97.09	465,822.17	426,375.85
C4.5	98.01	98.29	473,918.38	434,082.81
Oracle	100	100	479,583.91	438,923.30

station interest ranking

current system: free travel alerts – manually set up by traveller

future system: predict (and rank) the stations that travellers will visit in their future trips for personalised notifications

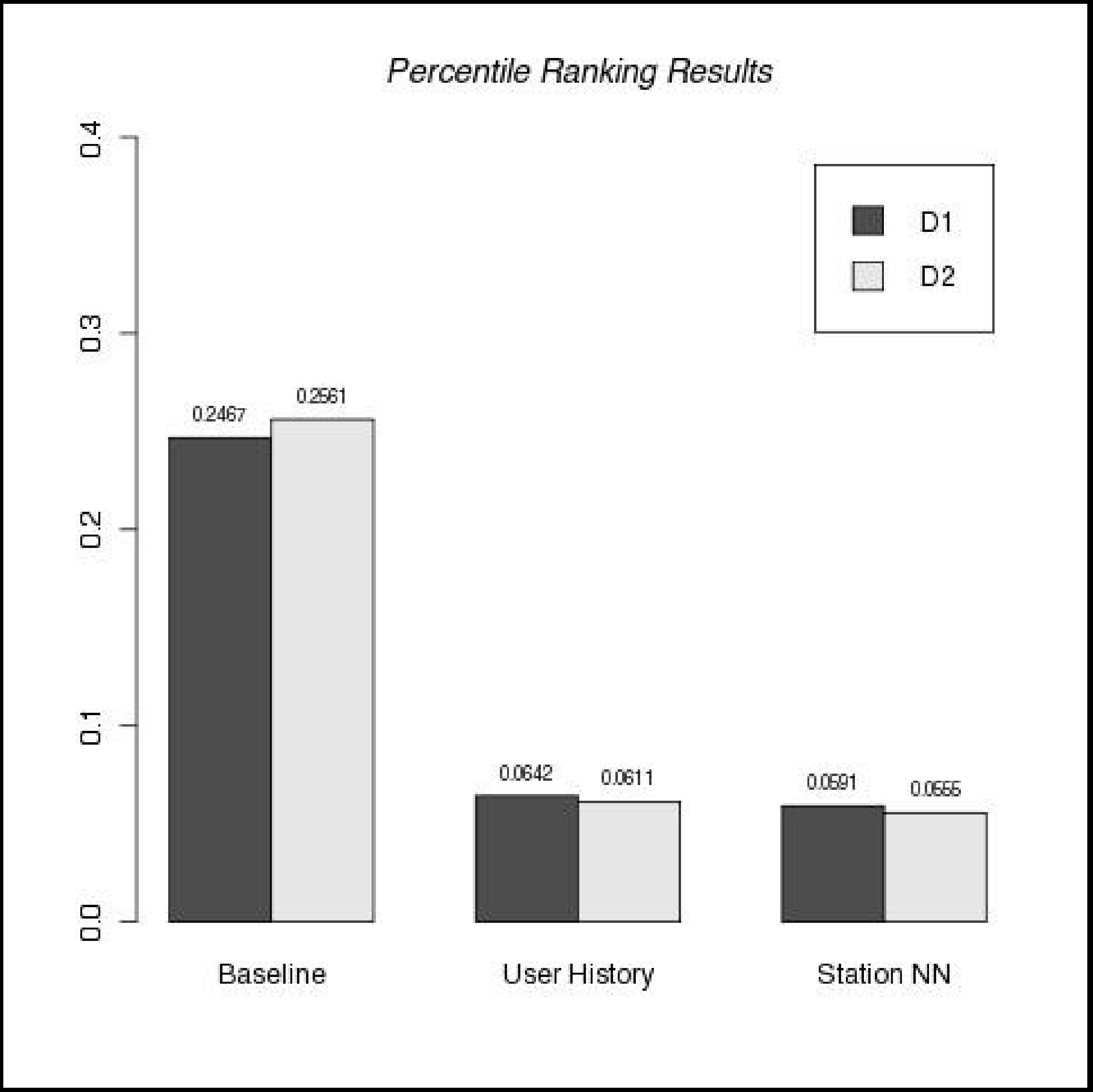
station interest ranking

can we automate this?

baseline: rank by visit popularity

proposal: station similarity neighbourhood (visit co-occurrence) and traveller trip history

station interest ranking



accurate ranking

without knowing who travellers are, the network topology, train schedule, disruptions and closures, we designed: **no context**

today:
(a) mobile location recommendations
(b) fare purchase recommendations
(c) travel alerts

Further reading:

D. Quercia, N. Lathia, F. Calabrese, G. Di Lorenzo, J. Crowcroft. **Recommending Social Events from Mobile Phone Location Data**. In IEEE ICDM 2010, Sydney, Australia.

N. Lathia, L. Capra. **Mining Mobility Data to Minimise Travellers' Spending on Public Transport**. In ACM KDD 2011, San Diego, USA.

N. Lathia, J. Froehlich, L. Capra. **Mining Public Transport Data for Personalised Intelligent Transport Systems**. In IEEE ICDM 2010, Sydney, Australia.

Android app to try: **www.tubestar.co.uk**

Questions?

@neal_lathia

neal.lathia@cl.cam.ac.uk