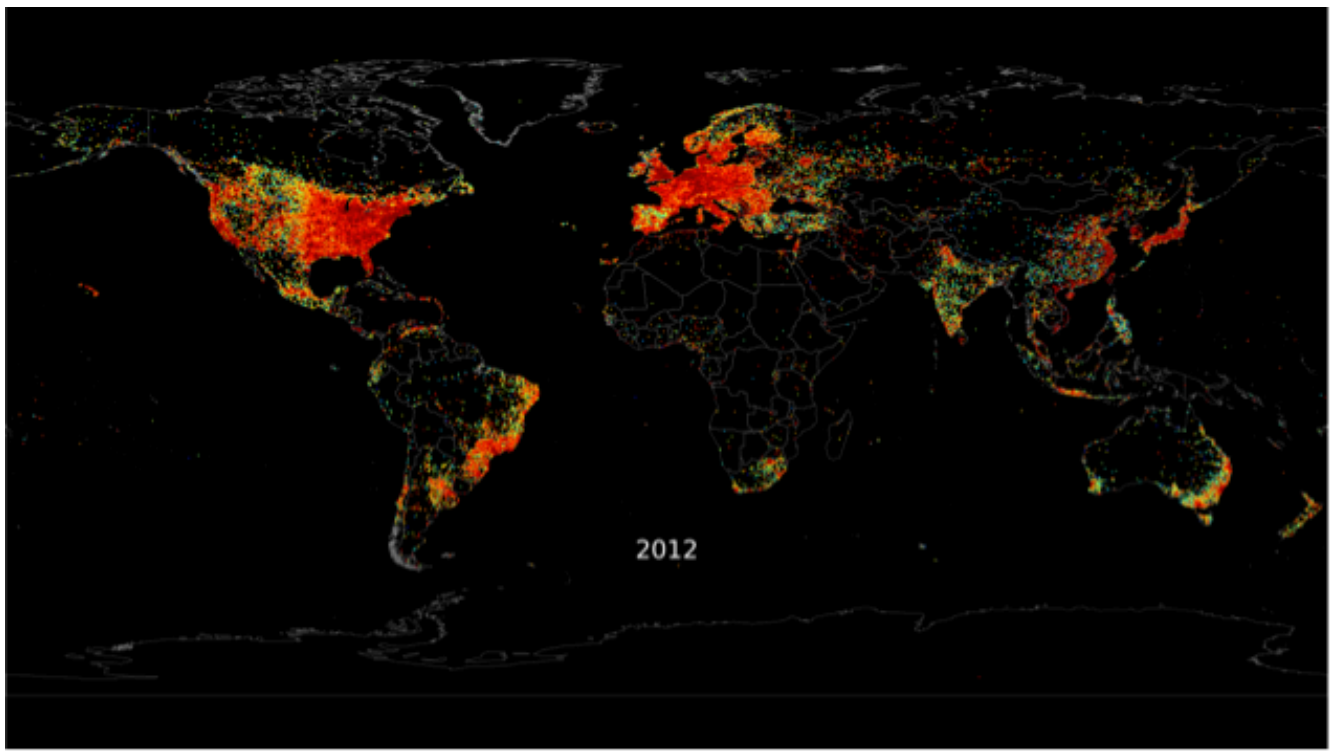
Economics of Technology

A trillion observations to infer social-economic behaviour



Klaus Ackermann
klaus.ackermann@monash.edu

Simon D. Angus

Paul Raschky

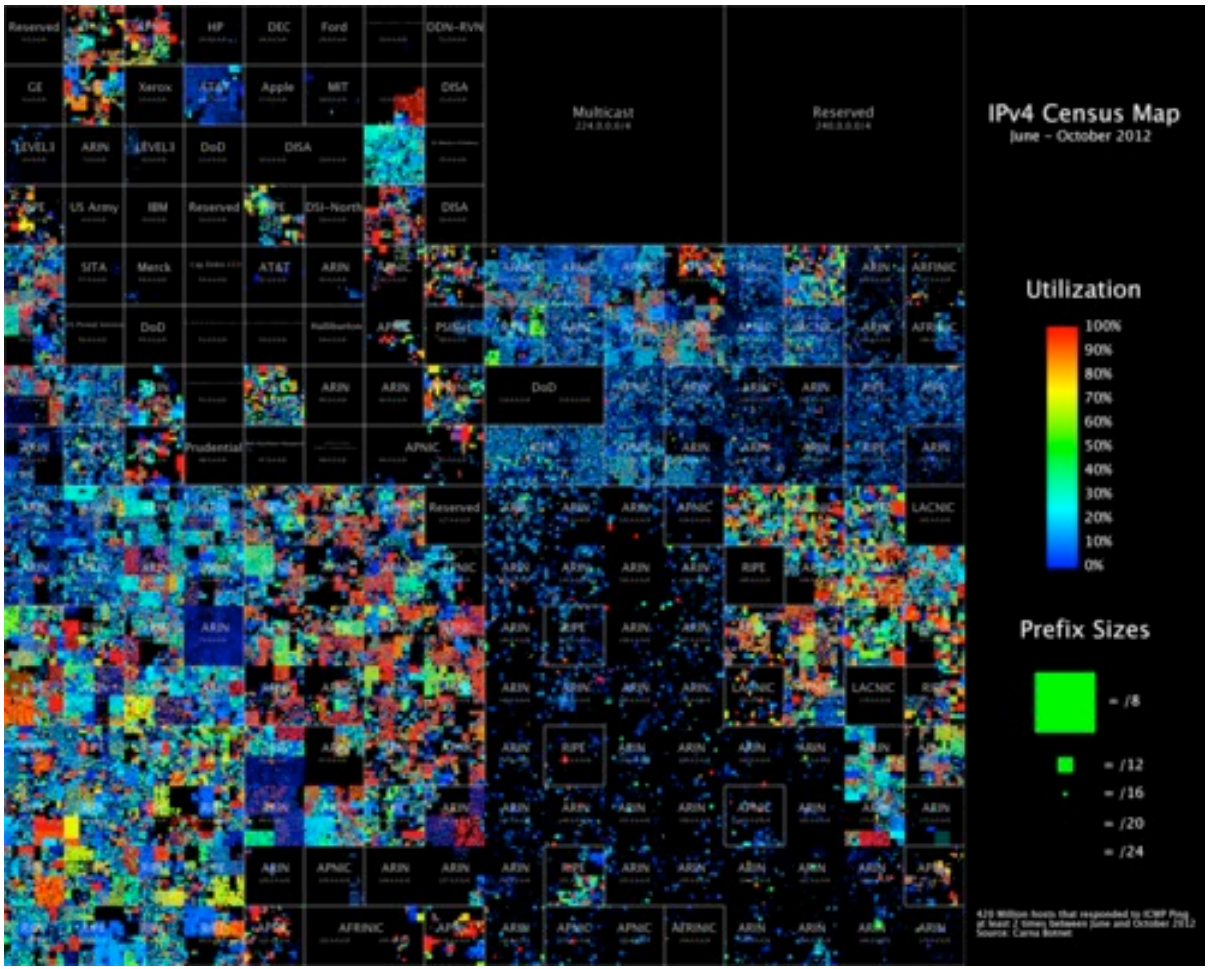
Department of Economics,
Monash Business School,
Monash University

MASSIVE
Multi-modal Australian ScienceS Imaging and Visualisation Environment

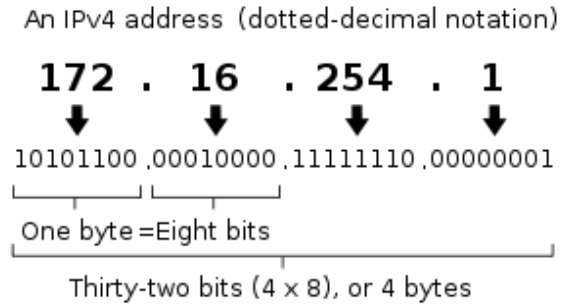
.au DOMAIN ADMINISTRATION LTD

MONASH University

Internet Protocol (IP) Addresses, IPv4, and Hilbert Projections



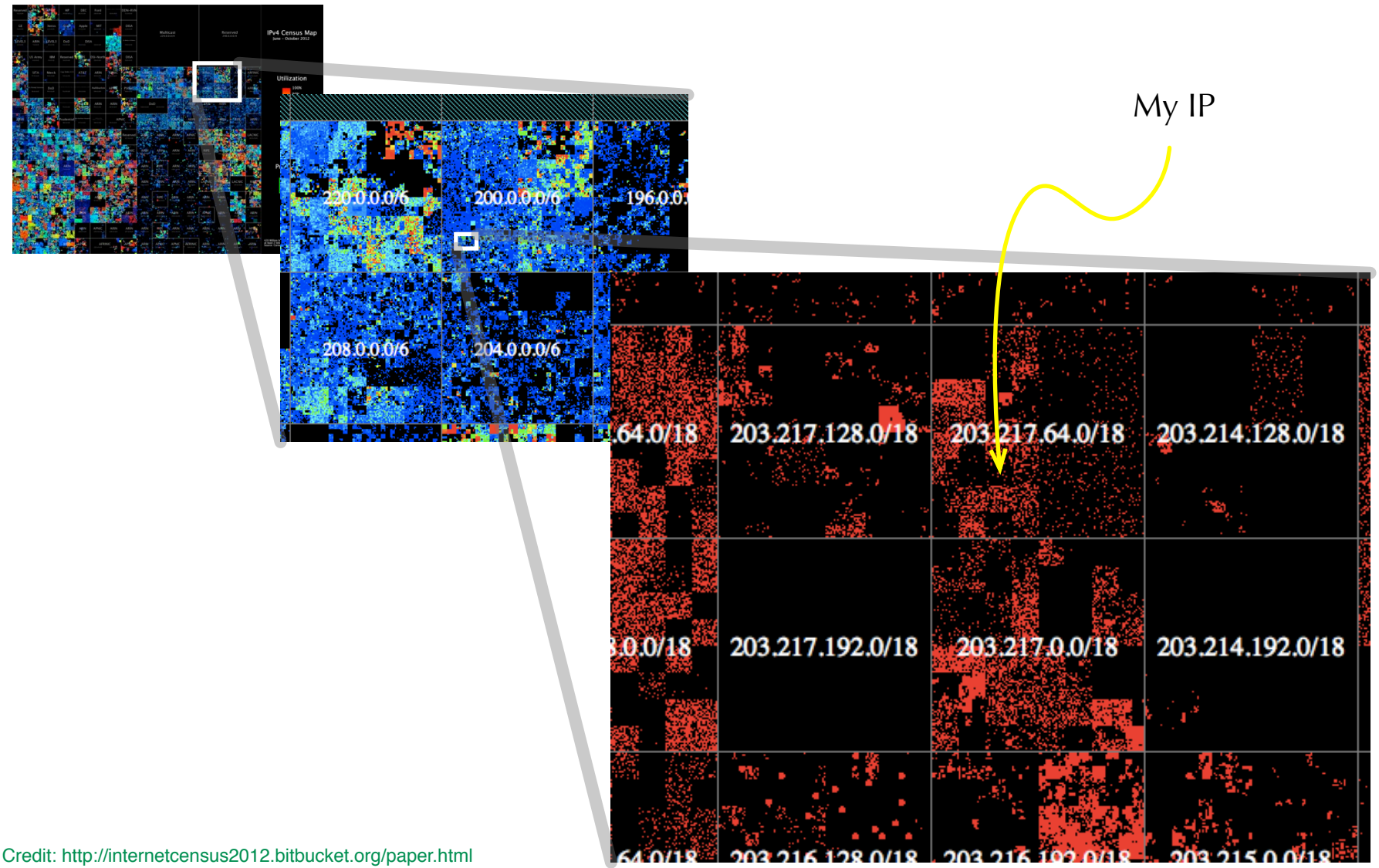
Credit: http://internetcensus2012.bitbucket.org/hilbert.html



Source: "Indeterminate' (via Wikimedia Commons)

Total possible:
4,294,967,296 (2^{32})
(> 4 billion)

Internet Protocol (IP) Addresses, IPv4, and Hilbert Projections



Credit: http://internetcensus2012.bitbucket.org/paper.html

The Idea

A Novel & Attractive Data Source …

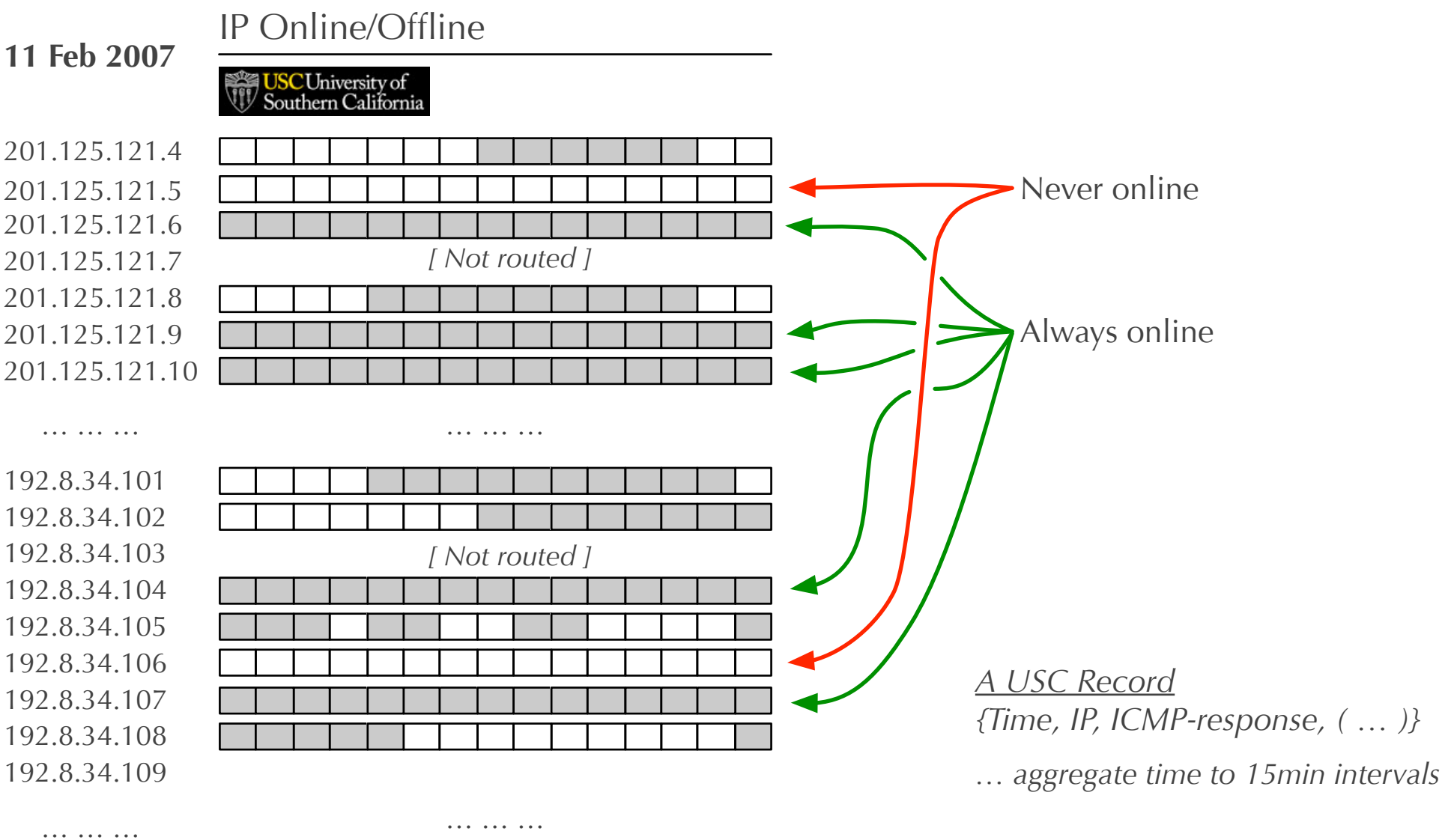
- Comprehensive: global, simultaneous, measurement (no border control for IP)
- Revealed vs. Stated: "*what* you do …" (not "what you *say* you do …").
- Granular: in time (intra-day) + space (Lat-Lon) (e.g. city-level).
- Accuracy: (limited) previous work uses poor location accuracy, here 10-40km.
- Date-range: 2005-2012 - critical time in internet's expansion.
- Diffusion of Technology: analysing the actual technology vs looking at records

Permitting Novel Social Science Questions …

- What are the main behavioural (sleep-wake, work-leisure) patterns of humankind (intra-day, inter-day, seasonal)?
- How has the diffusion of the internet affected democratic outcomes (at ballot-box level? in quasi-democratic countries?)
- Can internet activity reveal economic time-allocation?
- How affected by cultural norms is internet activity: religion?
- And so on …

The Data: USC, Digital Envoy .. to (IP-activity|time|geo-location)

11 Feb 2007



The Data: USC, Digital Envoy .. to (IP-activity|time|geo-location)

11 Feb 2007

IP Online/Offline

USC University of Southern California

201.125.121.4
201.125.121.5
201.125.121.6
201.125.121.7 *[Not routed]*
201.125.121.8
201.125.121.9
201.125.121.10

… … … … … …

192.8.34.101
192.8.34.102
192.8.34.103 *[Not routed]*
192.8.34.104
192.8.34.105
192.8.34.106
192.8.34.107
192.8.34.108
192.8.34.109

… … … … … …

IP —> Location

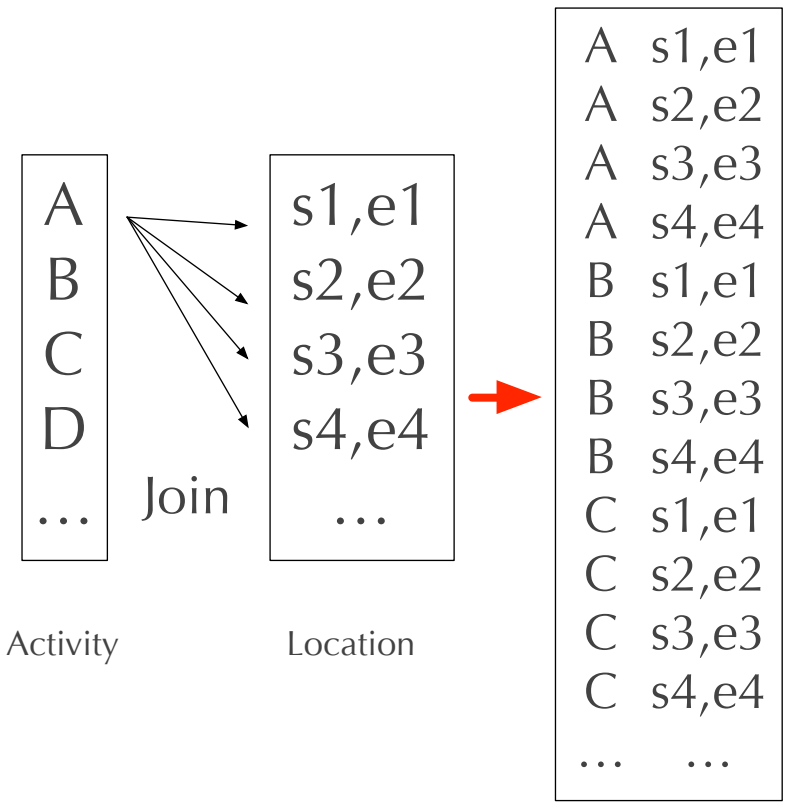
digital envoy

2007.Revision_k

A DE Record
{Time, IP-range, Lat, Lon, (…)}

Data joining & Processing

Normal join infeasible ...:
1.5 x 10^12 USC records
4 x10^11 DE records
.. ~ 6 x10^23 (600 sextillion records)

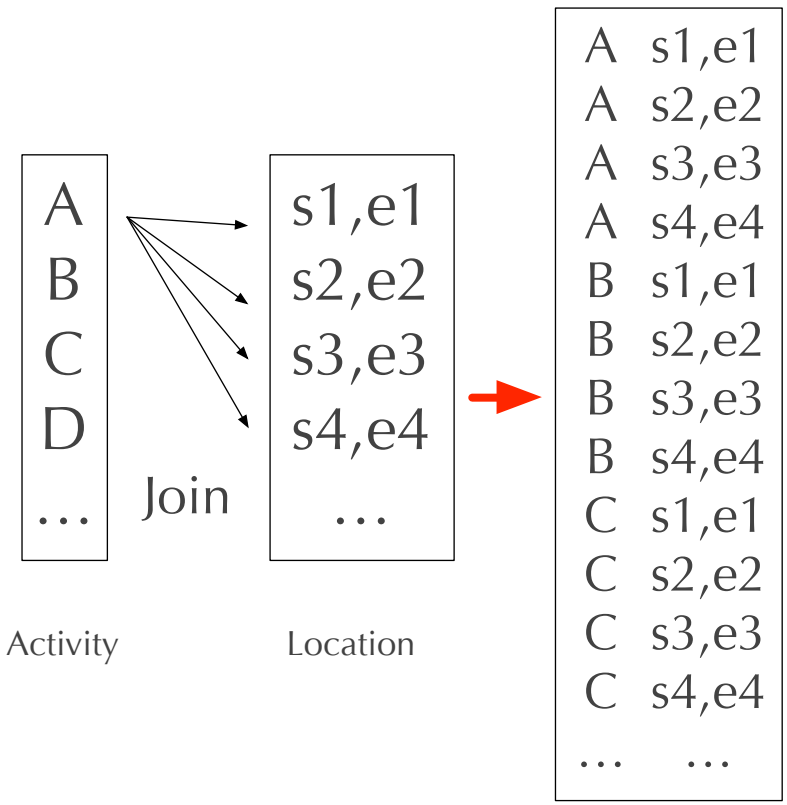


Standard solution: SQL Cartesian Product

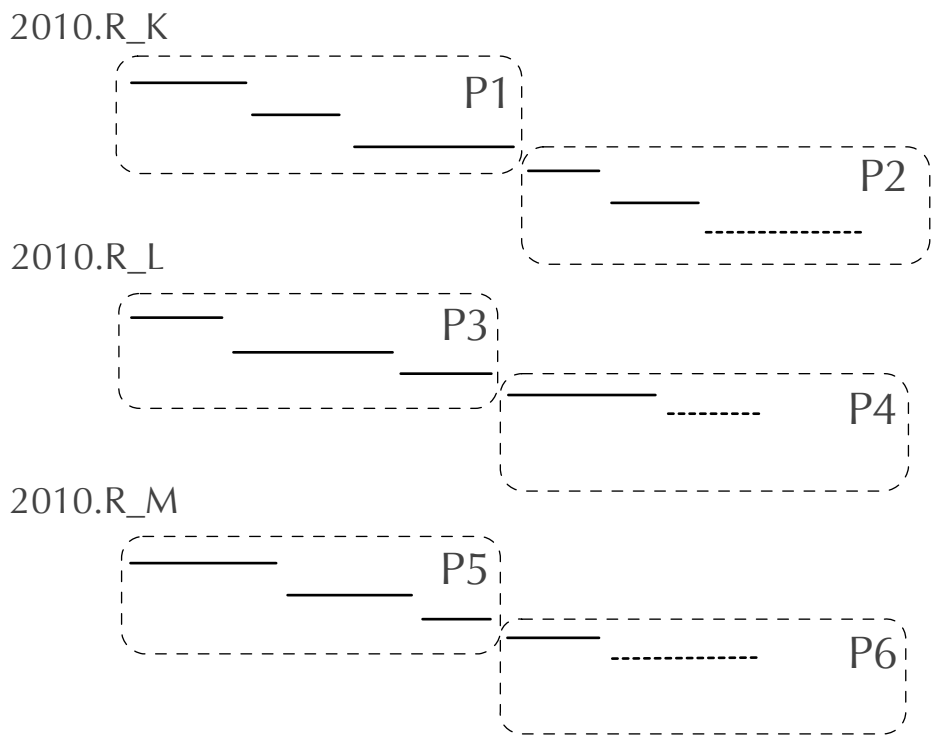
```
SELECT
        de.latitude,
        de.longitude,
        (u.timestamp div 900) as timeagregate,
        de.de_timestamp,
        SUM(if(u.on_off = 1, 1, 0)) as online,
        SUM(if(u.on_off = 0, 1, 0)) as offline
FROM
        usc AS u JOIN digitalenvoy de ON
        (u.probe_addr BETWEEN de.start_num AND de.end_num)
        and de.de_timestamp=(
                SELECT
                        dig.de_timestamp
                FROM
                        digitalenvoy dig
                WHERE
                        u.timestamp < dig.de_timestamp
                GROUP BY
                        dig.de_timestamp
                ORDER BY
                        dig.de_timestamp
                LIMIT 1)
GROUP BY
        de.latitude,
        de.longitude,
        timeagregate,
        de.de_timestamp
```

Data joining & Processing

Normal join infeasible ...:
1.5 x 10^12 USC records
4 x10^11 DE records
.. ~ 6 x10^23 (600 sextillion records)

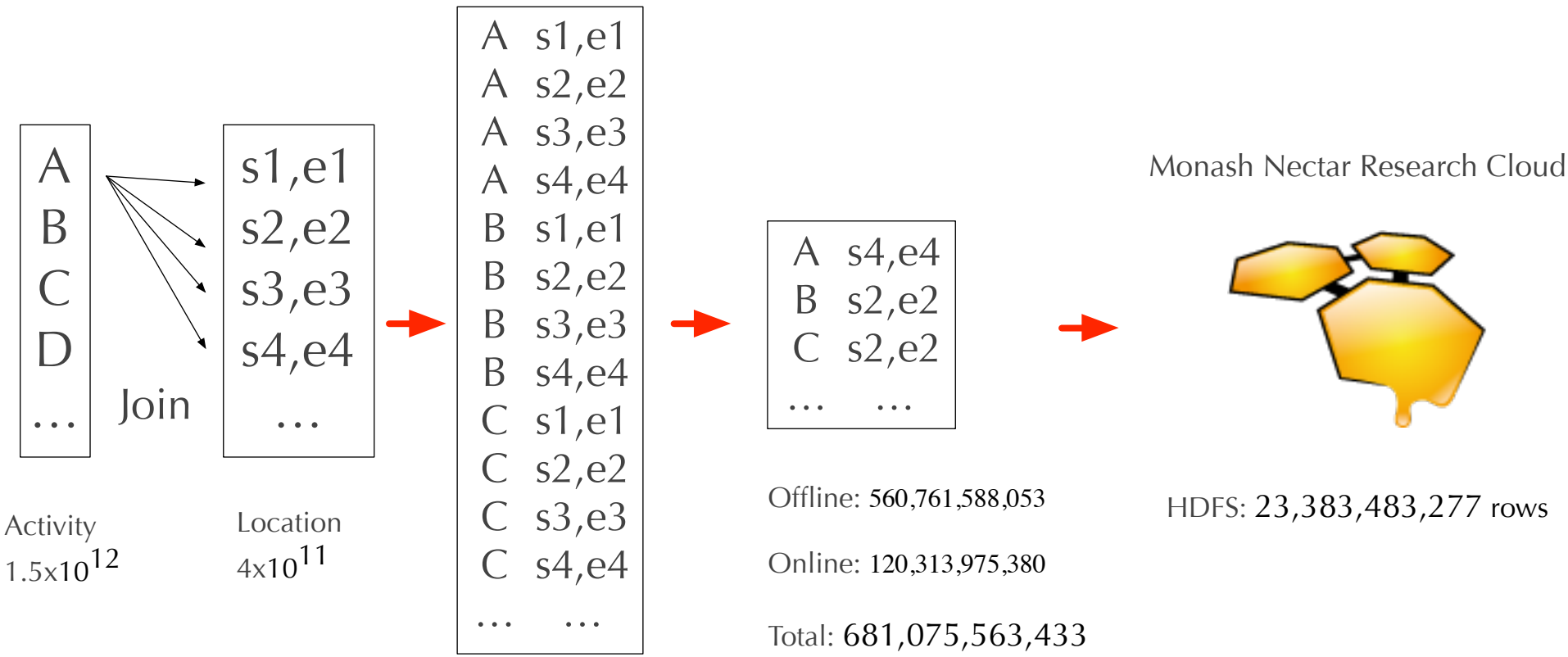


Our Approach: (effectively) index the Location (by range) DB, using a modified quantile algorithm, creating a look-up table by DB revision date and merging both lists with a runtime of approximate 2n in parallel



Data joining & Processing: Summary

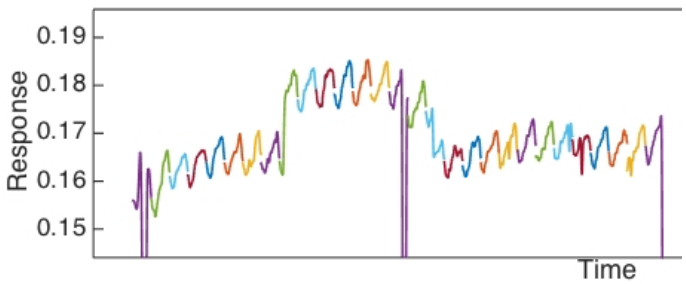
CPU hours: ~50000h = 5.7 years on one core

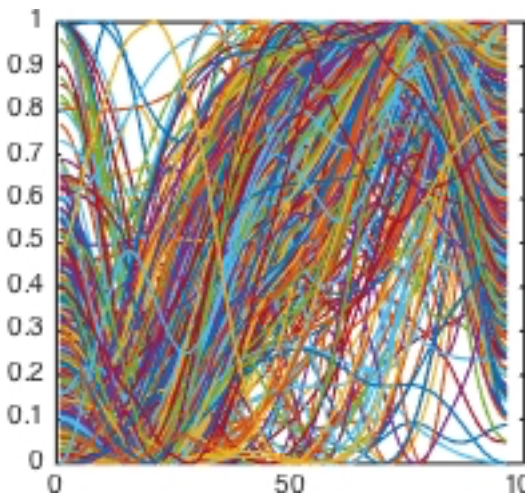


Processing Time: ~8 Month (Limited slots with enough RAM Synchrotron)

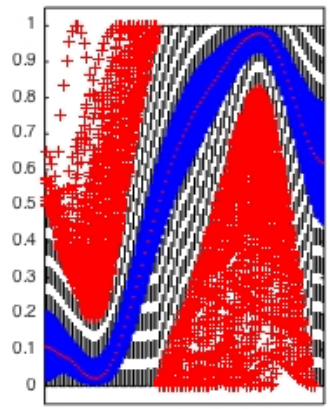
Aggregation Time: ~2h

From Raw to Useful: Example, London 2005-2011

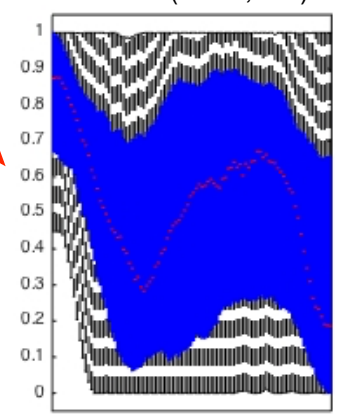




'Signal' (n=1,096, 92%)

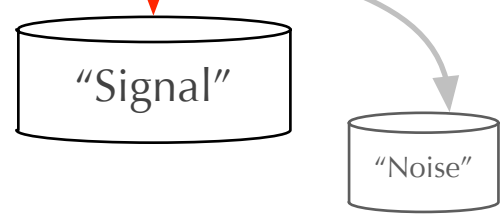


'Noise' (n= 90, 8%)



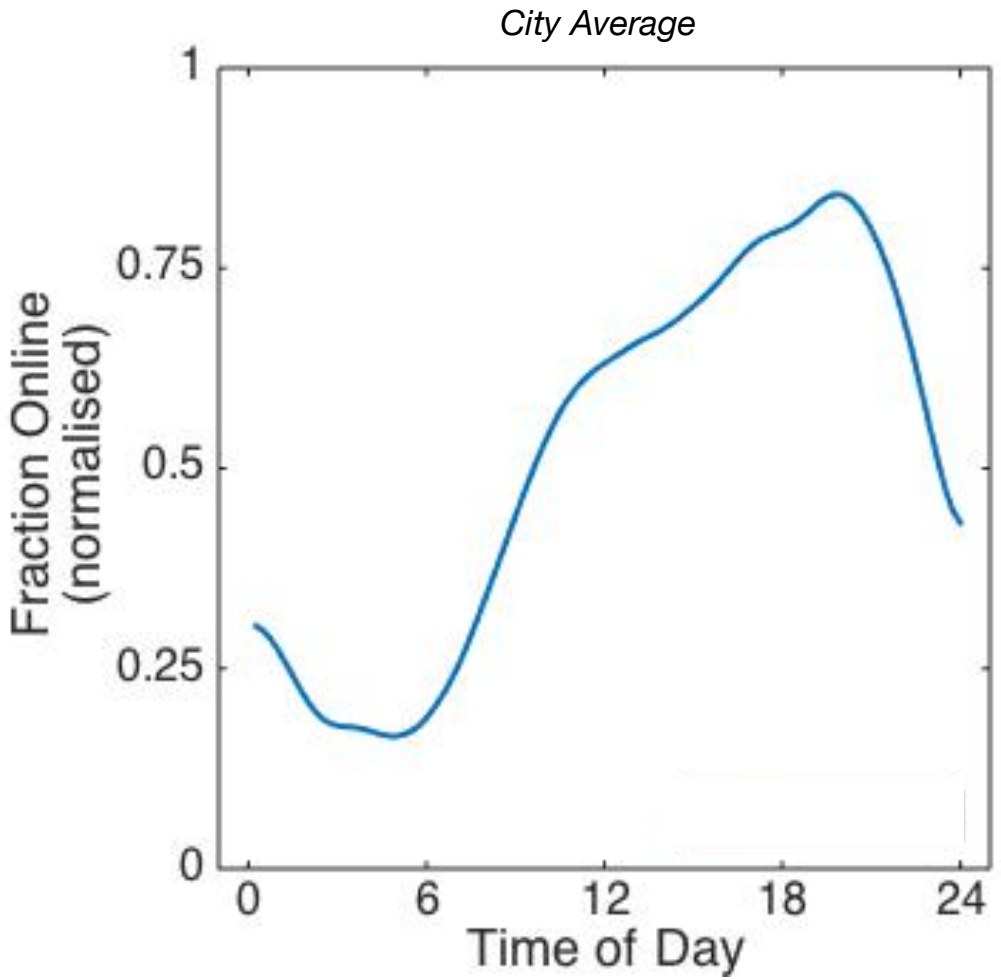
Single City Module

- Pre-filter (min online)
- Fraction_Online
- Cut by 24h, Daily Periods
- Robust Smooth, Normalise
- Multi-signal 1D Wavelet Decomposition
- Signal/Noise clustering

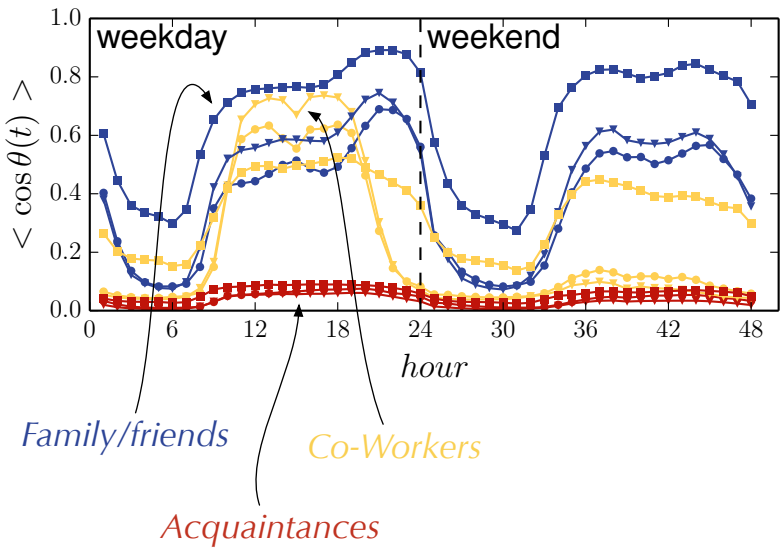


Data: London 2005-2011, raw traces (days): 1,539; filtered: 1,186 traces (days) (min 100 online per 15min)
Details: Clustering 'ward' (on Euclidean) of Wavelet analysis (sym3,lv6,coefs), Cophenetic Correlation: 0.9193

Anatomy of an intra-day trace

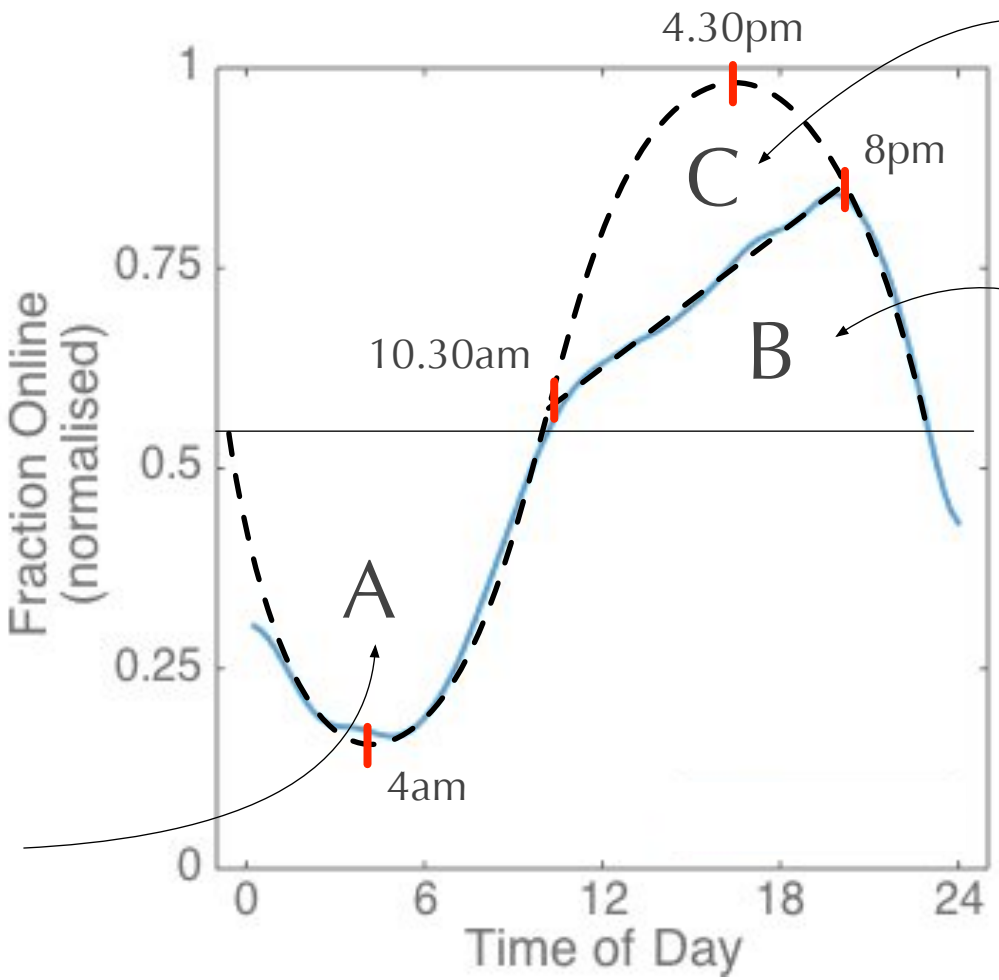


Toole et al (2015), "Coupling Human Mobility and Social Ties", arXiv: 1502.00690v1



Data: London 2005-2011, filtered + 'signal' only: 1,096 days (15 Dec 2005 .. 29 Dec 2011)

Anatomy of an intra-day trace



Substitution effect (away from personal IP use)

x

Personal day-time use effect

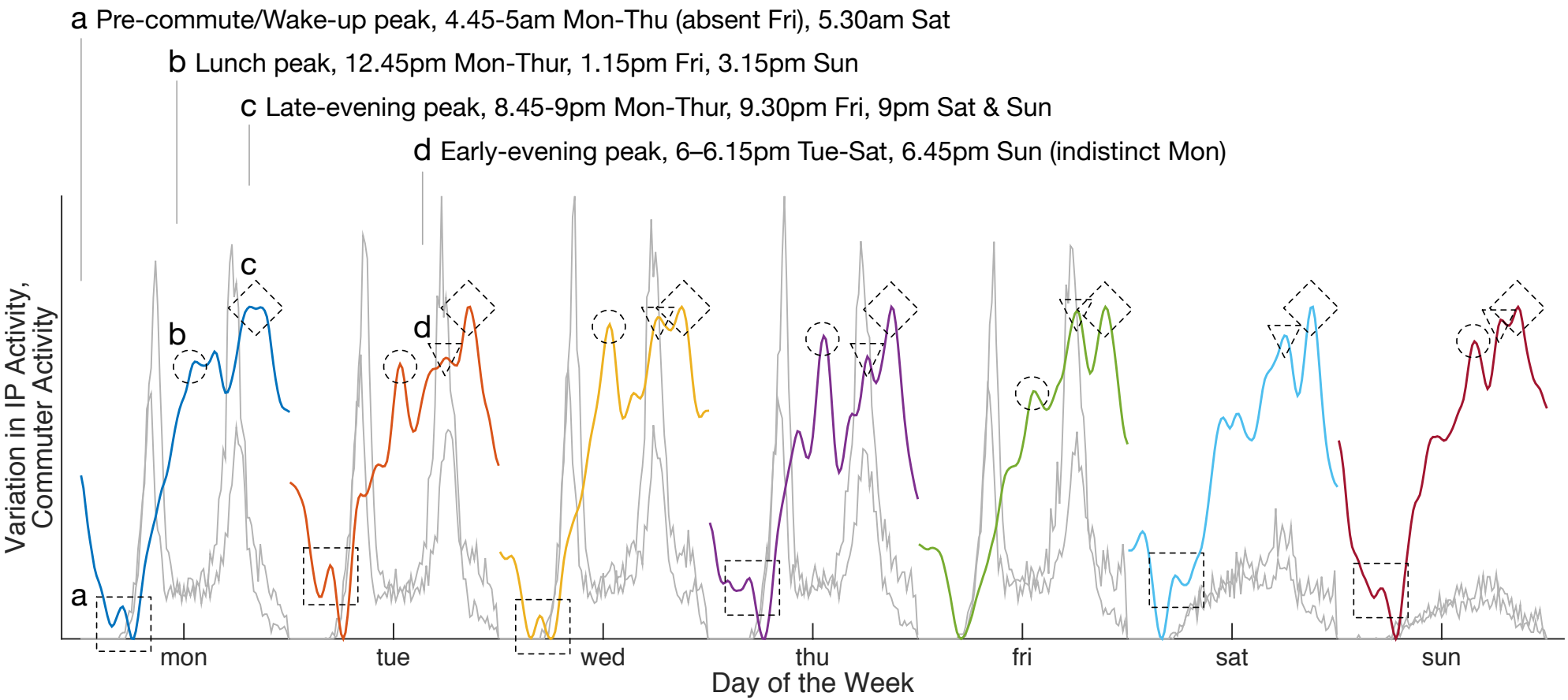
x

Sleep effect

A < B+C
(active-hours)

Data: London 2005-2011, filtered + 'signal' only: 1,096 days (15 Dec 2005 .. 29 Dec 2011)

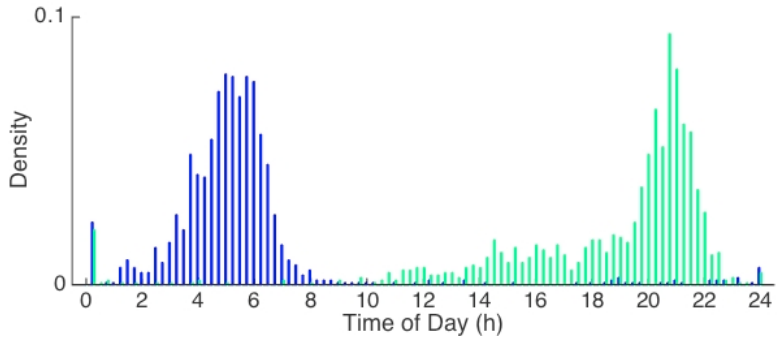
Daily IP Activity & Oyster-Card Intensity, London, GB



Oyster Activity: data from 5% sample of Oyster touch-on/touch-off activity restricted to LUL (LDN Underground) and NR (National Rail) events, two traces show 'inbound' and 'outbound' touch events

IP Data: data from 2 sets of contiguous months (Jun-Aug) in each year 2009, 2010; 126 days of data in all

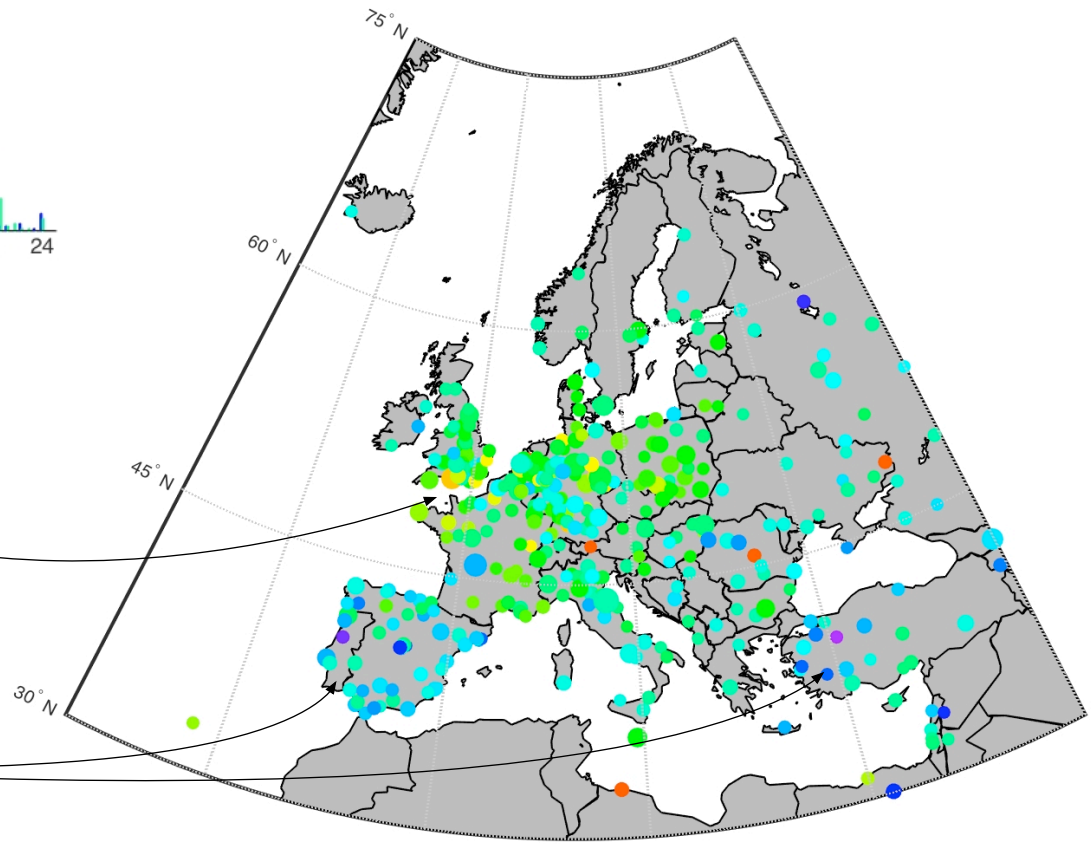
Multi-City Analysis: Time of Peak/Trough

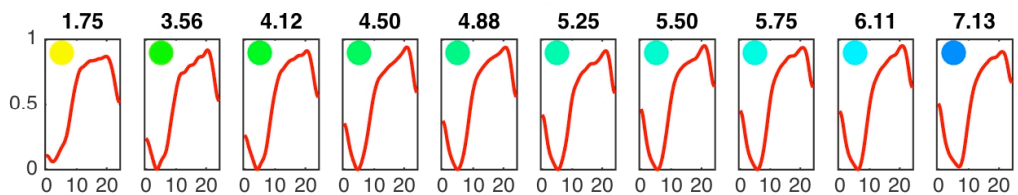


Time of Trough (24h)

Chanel Cities have earlier troughs

Spanish, Portuguese, and Turkish cities have later trough

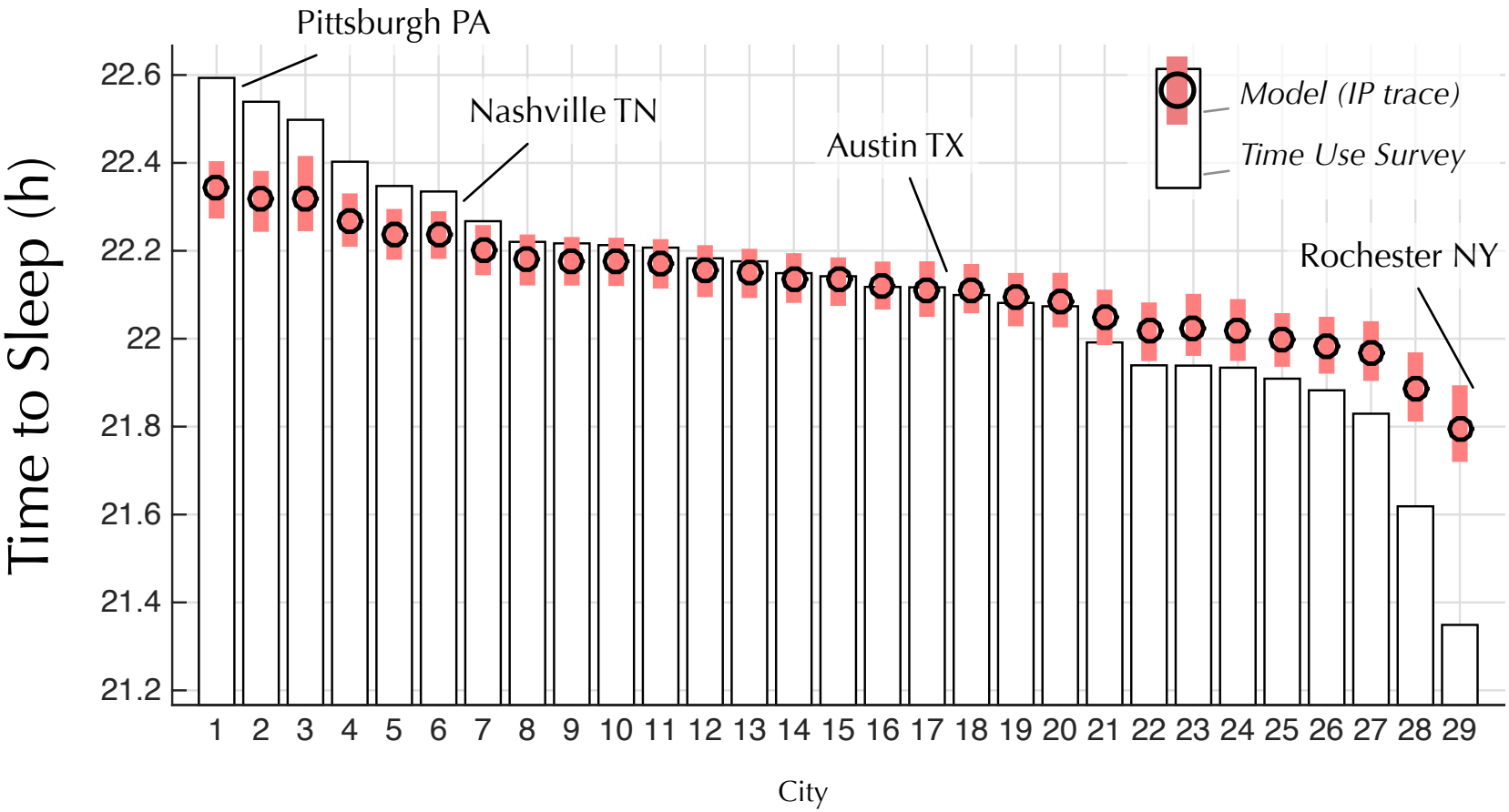




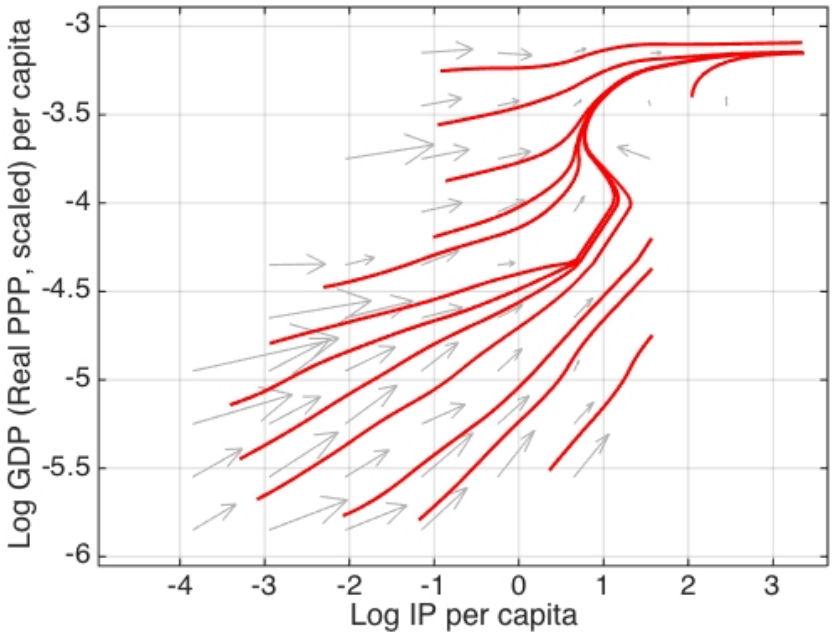
Data: 1,065 cities after pre-filtering and processing.

American Time Use Survey: Up-Scaling of a traditional survey

- Use the internet data as an empirical proxy for human behaviour at a very fine temporal and spatial scale
- Idea: Find a model that predicts the start and end sleep and work times based on the shape of the internet trace by Metropolitan Statistical Areas (MSA) in the US



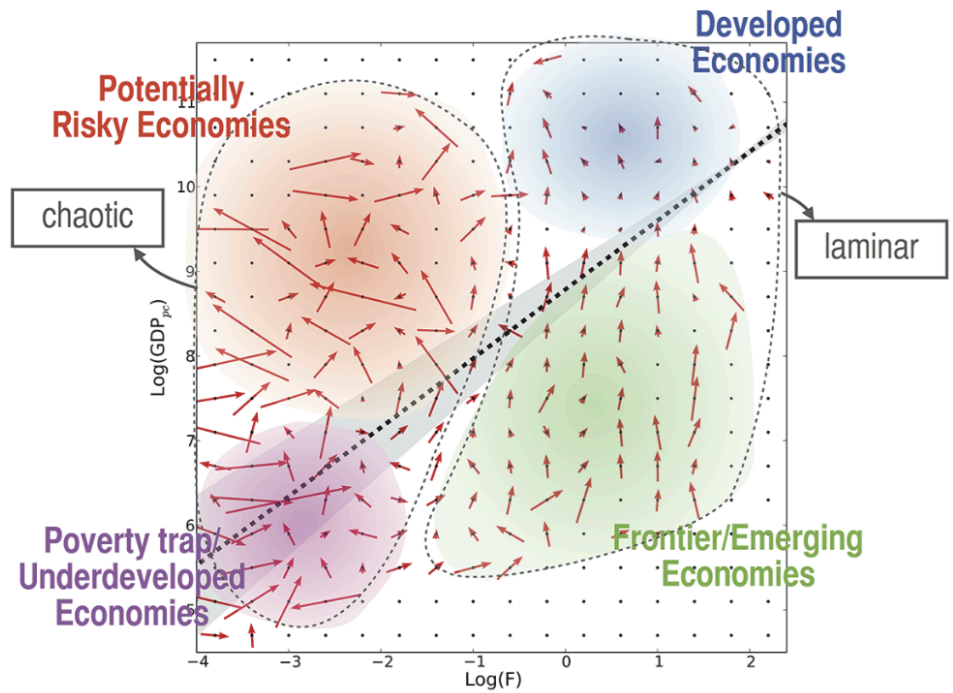
The S-Curve of Technological Diffusion



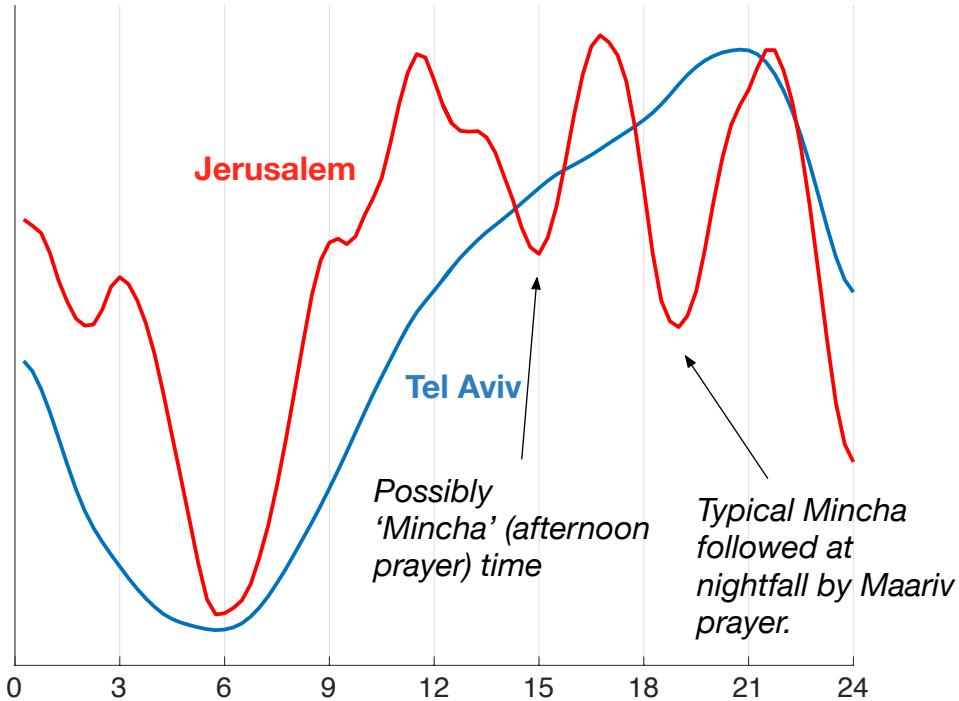
Cristelli, M., Tacchella, A., & Pietronero, L. (2015). The Heterogeneous Dynamics of Economic Complexity. *PLoS ONE*, *10*(2), e0117174–15

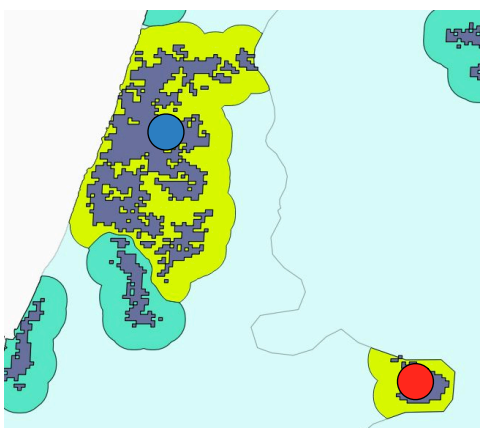
GDP City Level:

- Based on OECD regional accounts TL2 and TL3 rescaled using Landsat 2006 population raster GIS data and NYU metropolitan blocks
- Real GDP PPP city level (left)
- Nominal GDP PPP country level (right)



Religion: Revealed vs Stated Preferences

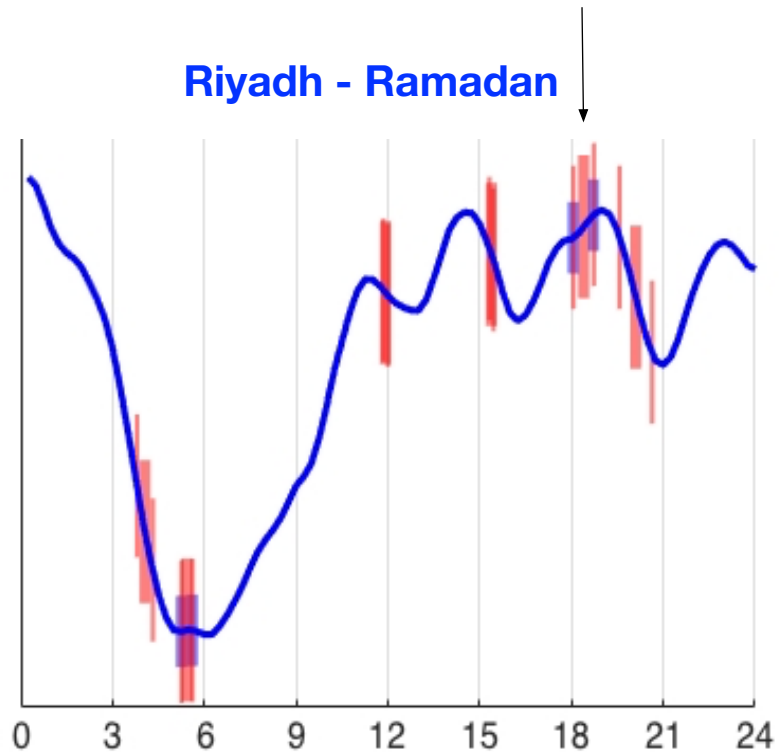




Urban blocks (2000) & the buffered area

Different Prayer times in different Religions

For Suni the fast can be broken at the start of the 5th prayer



Discussion

So far

- Successful handling, conversion & cleaning of trillions of IP-activity observations, linked to accurate geo-location
- Successful preliminary analysis tools developed on basic and more complex properties of ip-activity

Preliminary Observations

- Strong spatial-correlation of ip-activity traces, e.g. Oyster and Sleep
- Good evidence of discontinuities at political boundaries suggesting cultural/ institutional factors driving behaviour

Current Work & Future

- Publication of the Data-Set for Australia as well as the cities world wide
- Internet censorship and political elections with evidence from Russia
- Contact me: klaus.ackermann@monash.edu