

An agile approach to knowledge discovery of web log data

Paul Lam, Thibaut Sacreste, Paul Ingles

OR54, Edinburgh, 4 September 2012

Why web log data

Visitor information

- ❖ web page requested
- ❖ client IP address
- ❖ request timestamp
- ❖ query string
- ❖ bytes served
- ❖ user agent
- ❖ referrer

uSwitch

- ❖ an online business
- ❖ 100 GB of uncompressed data per month

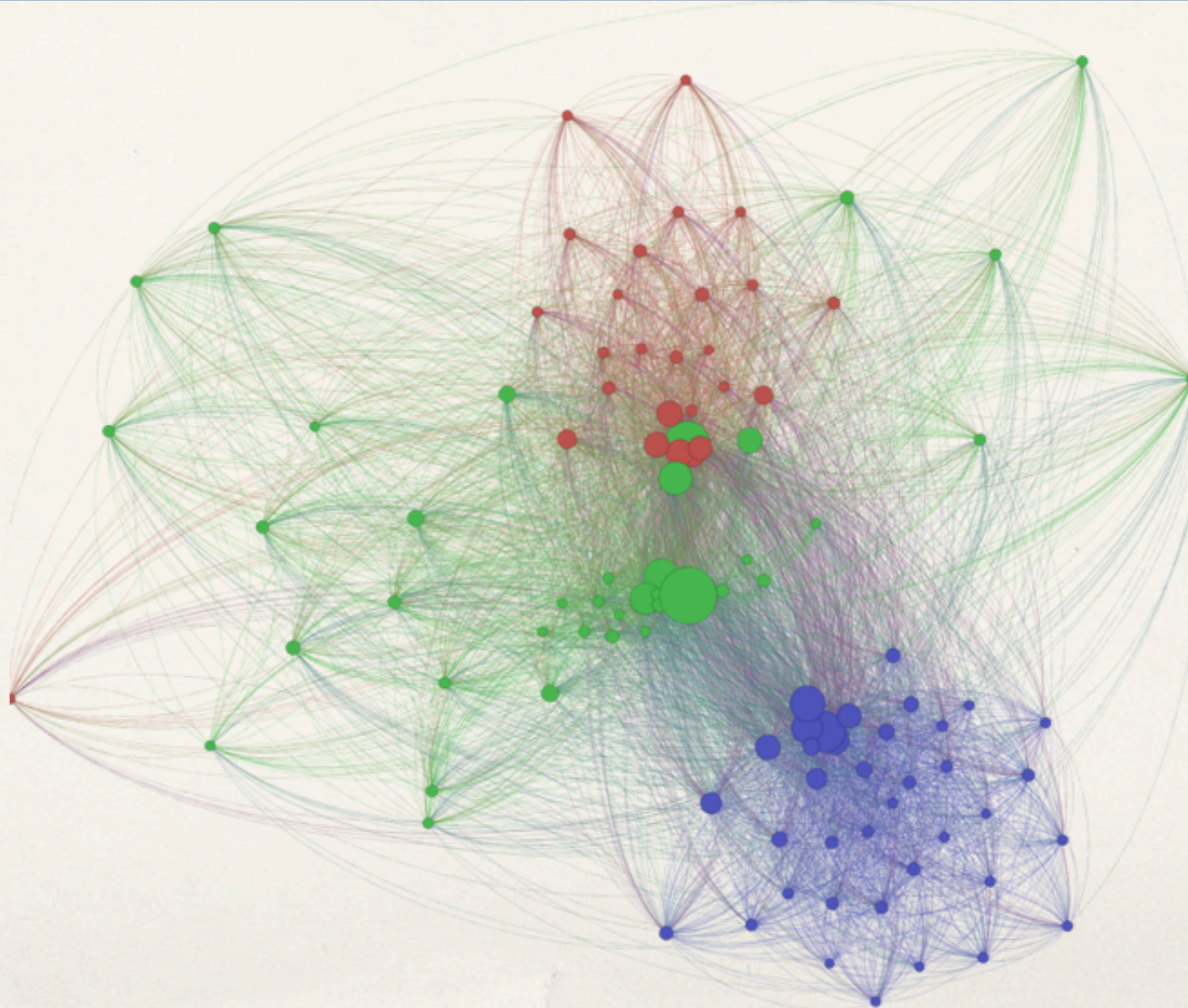
The screenshot shows the uSwitch website homepage. At the top, there is a navigation bar with the uSwitch logo and links for "Gas, electricity & water", "Broadband, TV & home phone", "Mobiles", "Money & insurance", and "About". Below the navigation bar, a welcome message reads: "Welcome back! We've remembered your postcode is NW1 0DU. [Not you?](#)". The main headline is "Join the rest of London and save £1,000*". Below the headline, there are three main service categories, each with an image and a call-to-action button:

- Save up to £420 on energy bills***: Image shows a mug with logos for British Gas, E.ON, SSE, EDF, and ScottishPower. Call-to-action: "Switch your gas & electricity".
- Broadband deals from £2.50**: Image shows a laptop displaying logos for Virgin, TalkTalk, BT, and Sky. Call-to-action: "Compare broadband deals".
- Switch to a 0% Credit Card**: Image shows several credit cards. Call-to-action: "Compare credit cards".

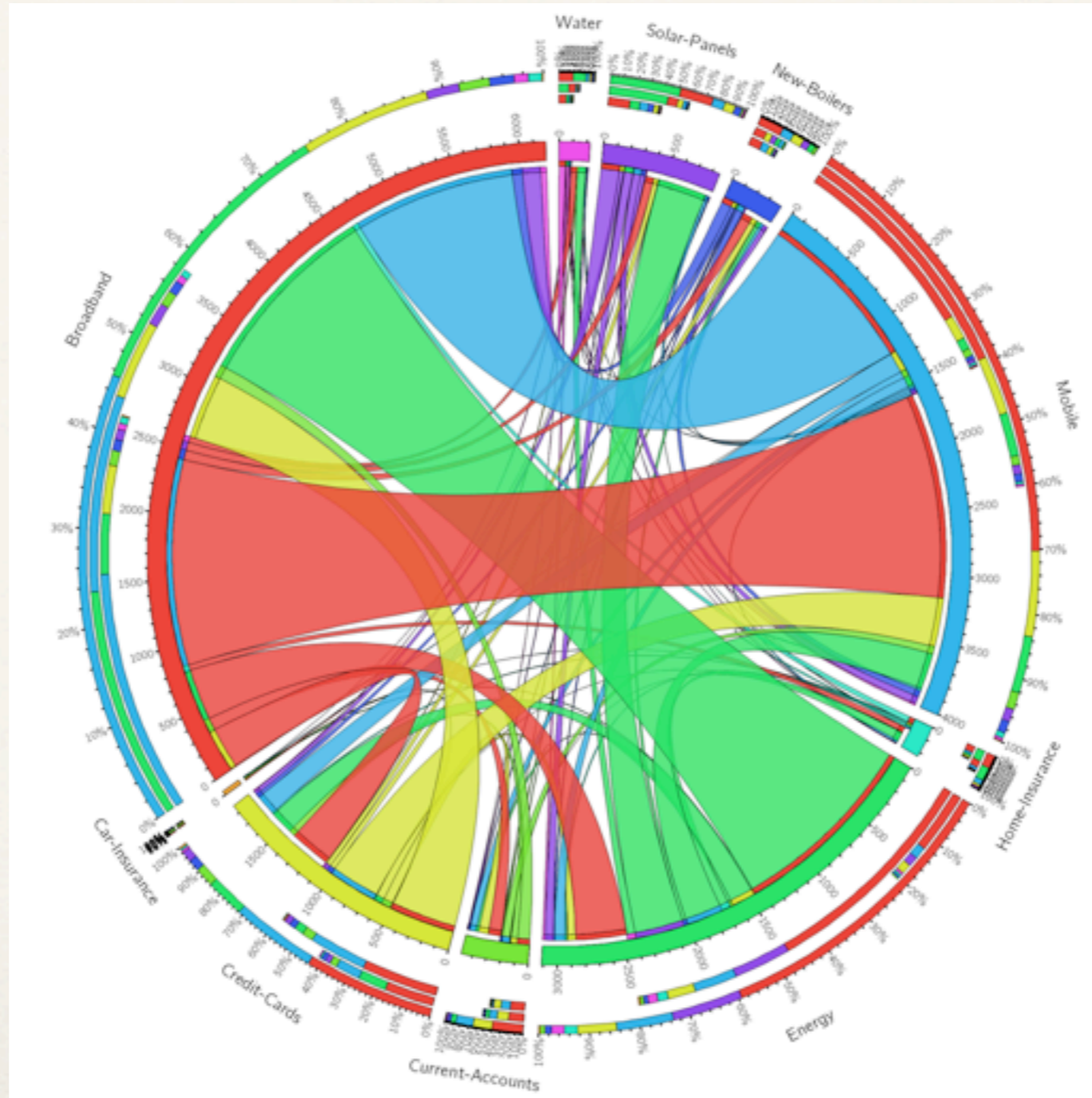
Each category also has a list of sub-links below the main button:

- Under "Switch your gas & electricity": Gas & electricity, Free solar panels, Boiler cover.
- Under "Compare broadband deals": Broadband deals, Broadband & home phone, Broadband & TV packages.
- Under "Compare credit cards": 0% balance transfer cards, 0% purchase cards, Credit builder cards.

Behavioural analysis

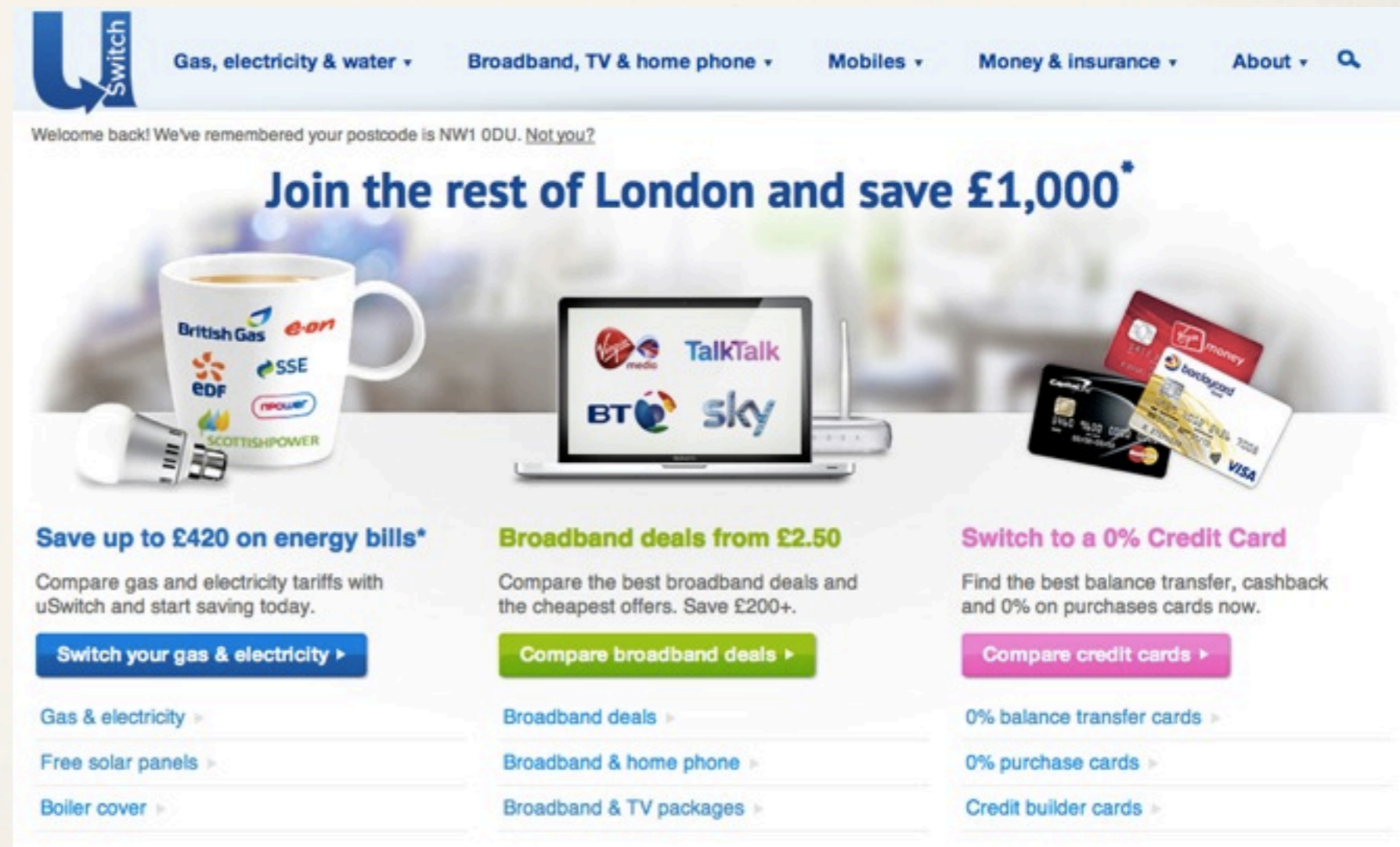


Purchasing habits



Product personalisation

- ❖ 30% of Amazon sales comes from its recommendation engine [1]
- ❖ Examples on uSwitch homepage



The screenshot shows the uSwitch homepage with a personalized message: "Welcome back! We've remembered your postcode is NW1 0DU. [Not you?](#)" Below this, a large banner reads "Join the rest of London and save £1,000*". The banner features three main categories of offers:

- Save up to £420 on energy bills***: Accompanied by an image of a mug with logos for British Gas, e-on, EDF, SSE, and SCOTTISHPOWER. A button says "Switch your gas & electricity >". Below are links for "Gas & electricity >", "Free solar panels >", and "Boiler cover >".
- Broadband deals from £2.50**: Accompanied by an image of a laptop showing logos for Virgin mobile, TalkTalk, BT, and sky. A button says "Compare broadband deals >". Below are links for "Broadband deals >", "Broadband & home phone >", and "Broadband & TV packages >".
- Switch to a 0% Credit Card**: Accompanied by an image of several credit cards. A button says "Compare credit cards >". Below are links for "0% balance transfer cards >", "0% purchase cards >", and "Credit builder cards >".

[1] Schumpeter, "Building with big data", Economist, 26 May 2011

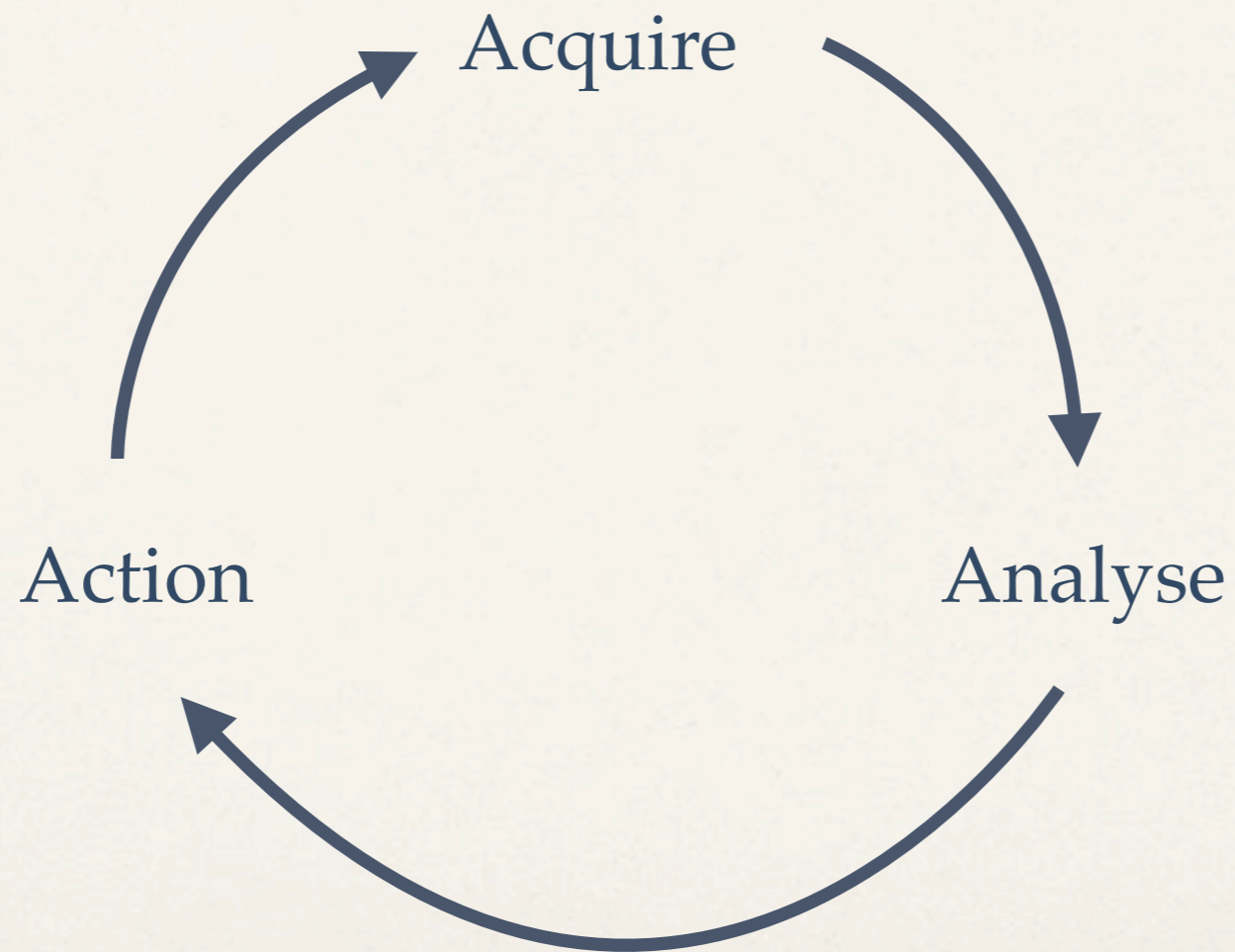
Goals

- ❖ Exploration of data
- ❖ Exploitation of data

Data team at uSwitch

- ❖ a core team of 3 complementary skilled people:
 - ❖ data scientist
 - ❖ back-end developer
 - ❖ software architect
- ❖ not a boundary of our roles
 - ❖ guess who loves ggplot and who does the NLP work
- ❖ collaborate with domain experts (designers, marketers, product managers, developers, etc) across the company

Challenges and Solutions

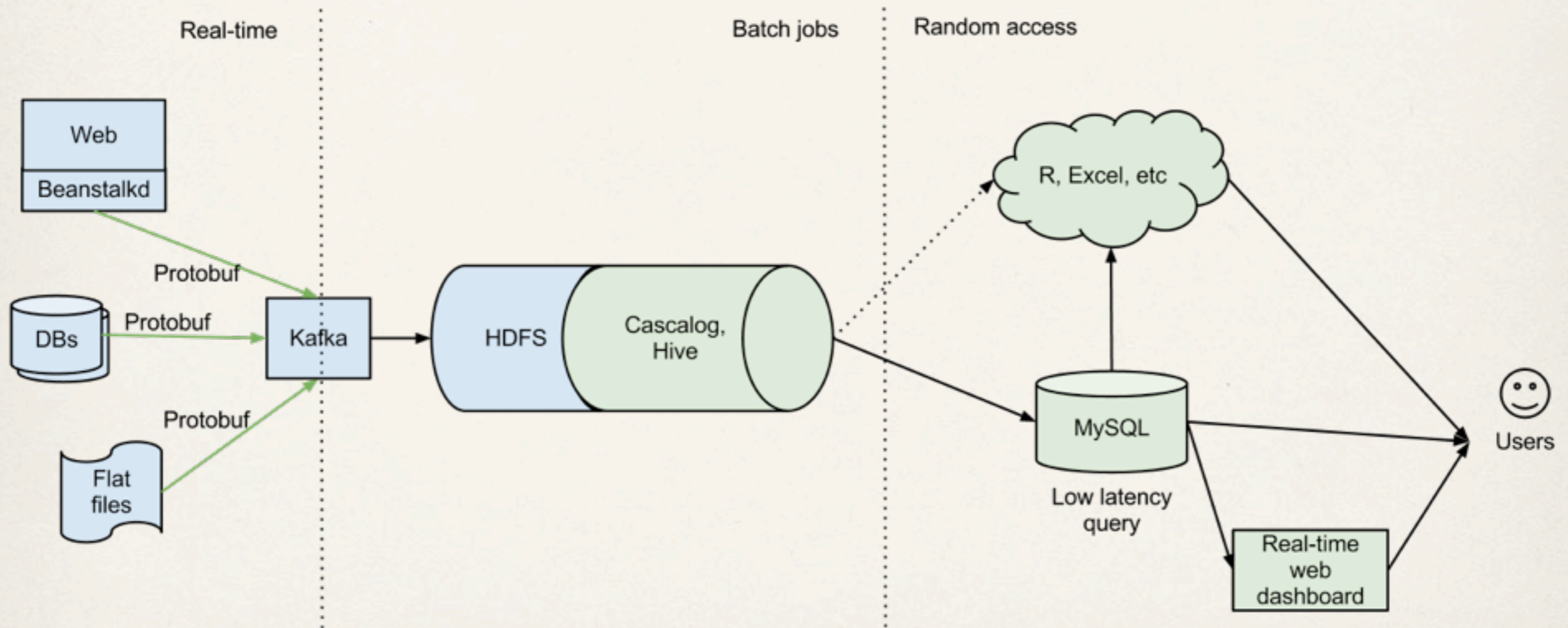


Acquire

Data extraction considerations

- ❖ hundreds of applications distributed over ~50 Amazon EC2 instances
- ❖ 10+ of the apps are actively worked on at any given time
- ❖ projects are owned by small, autonomous teams
- ❖ great for the business, not so great to get data from

Distributed data pipeline



Analyse

One of two millions a day

- * `{:status 200, :scheme http, :pipe ., :request-uri /broadband/?gclid=CPnYgdqj0bECFa4mtAodVEsAYA, :http-x-forwarded-for 92.9.200.50, :msec 1344196910.137, :sent-http-set-cookie -, :body-bytes-sent 18836, :query-string gclid=CPnYgdj0bECa4mtAdVEsAYA, :request-content-type -, :cookie-urefs -, :request GET /broadband/?gclid=CPnYgdj0bECa4mtAdVEsAYA HTTP/1.1, :upstream-response-time 0.164, :sent-http-content-type text/html, :hostname nginx-lb-20120229-1942-24.uswitchinternal.com, :sent-http-location -, :time-local 05 / Aug / 2012:20:01:50 +0000, :http-referer http://www.google.co.uk/aclk?sa=l&ai=D1556&rct=j&q=best%20value%20internet%20uk, :http-user-agent Mozilla/5.0 (Windows NT 6.0) AppleWebKit/537.1 (KHTML, like Gecko) Chrome/21.0.1180.60 Safari/537.1, :request-time 0.164, :request-body -, :http-host www.uswitch.com, :upstream-addr 178.32.60.100:80, :sent-http-server -, :upstream-status 200, :uscc <ANON>}`

Ad-hoc queries - Apache Hive

Editor Schema History Saved Queries Custom Functions Help

Autocomplete: Are you are struggling to remember the name of a table, column or function? try typing the first few characters and then hit **ESC** to see suggestions.

```
1
```

Submit Query

Recent queries for:

twitter conversion rate report

COMPLETED

Preview Download Edit Save Delete

Author	Submitted @	Started @	Completed @	Queued for	Query took	Records
Paul Lam	2012-08-30 17:21:20	2012-08-30 17:21:20	2012-08-30 17:25:28	less than a second	4 minutes	3

```
1 SELECT imp.year, imp.month, imp.referrer, imp.cnt AS impression_count, swt.cnt AS switch_count, ROUND((swt.cnt / imp.cnt * 100), 2) AS conversion_pct
2 FROM
3 (
4   SELECT YEAR(dated) AS year,
5          MONTH(dated) AS month,
6          CONCAT(PARSE_URL(http_referer, 'HOST'), PARSE_URL(http_referer, 'PATH')) AS referrer,
7          COUNT(1) AS cnt
8   FROM uswitch_data_sessionised_weblog
9   WHERE (
10          http_referer LIKE 'http://t.co/%'
11          OR http_referer LIKE 'http://bit.ly/%'
12        )
13  AND status = 200
```


Word Count - Cascalog

```
(defmapcatop split [line]
  "reads in a line of string and splits it by regex"
  (s/split line #"[\[\]\(\),.\s]+"))

(defn -main [in out & args]
  (?<- (hfs-delimited out)
    [?word ?count]
    ((hfs-delimited in :skip-header? true) _ ?line)
    (split ?line :> ?word)
    (c/count ?count)))
```


TF-IDF

- ❖ Extended from word count example
- ❖ Single-purpose methods
- ❖ Composition of functions
- ❖ github.com/Quantisan/Impatient
- ❖ github.com/Cascading/Impatient

```
(defn D [src]
  (let [distinct-doc-id (uniquefy (select-fields src ["?doc-id"]))]
    (<- [?n-docs]
      (distinct-doc-id ?doc-id)
      (c/count ?n-docs))))

(defn DF [src]
  (let [distincted (uniquefy src)]
    (<- [?df-word ?df-count]
      (distincted _ ?df-word)
      (c/count ?df-count))))

(defn TF [src]
  (<- [?doc-id ?tf-word ?tf-count]
    (src ?doc-id ?tf-word)
    (c/count ?tf-count)))

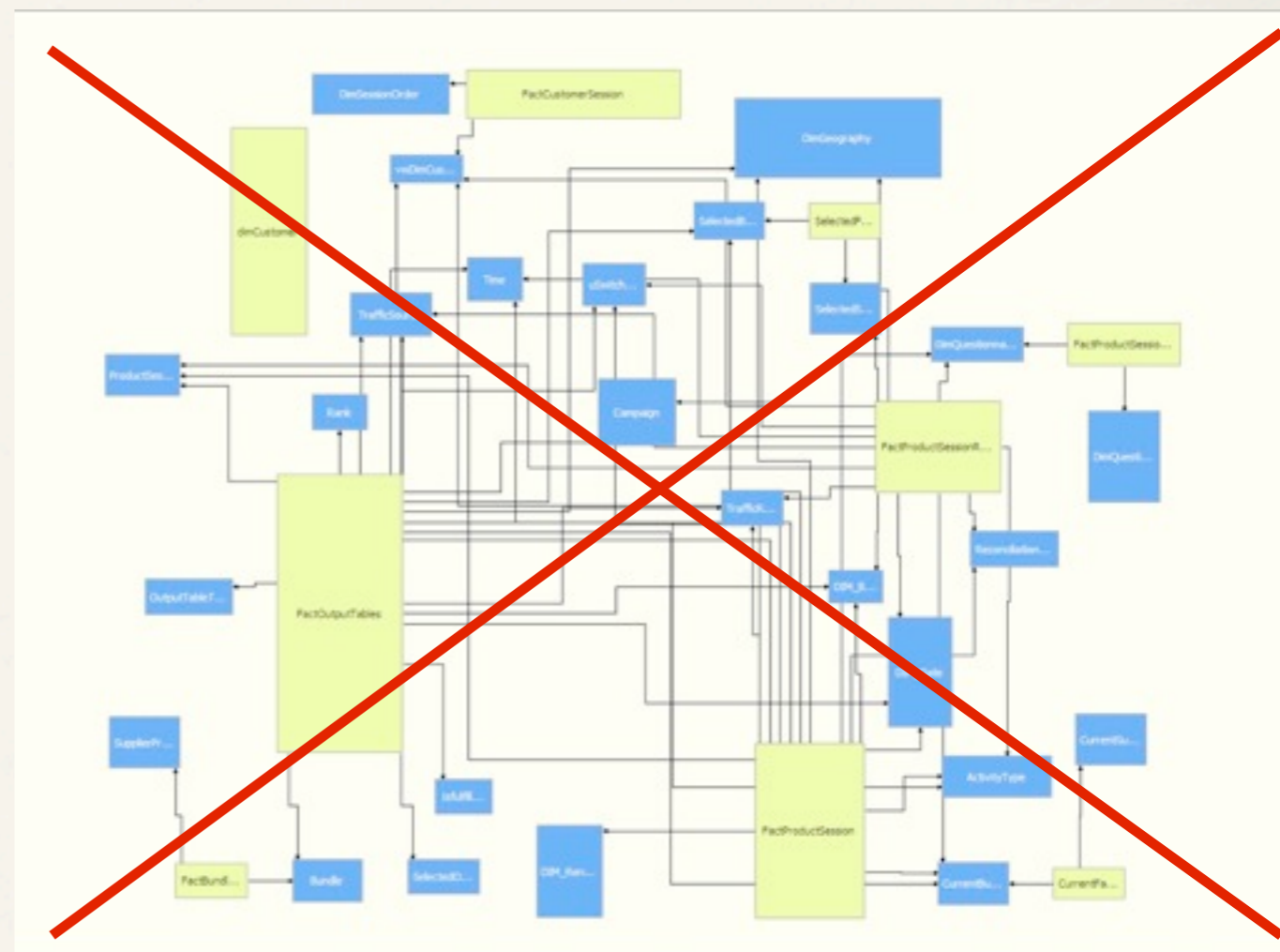
(defn tf-idf-formula [tf-count df-count n-docs]
  (->> (+ 1.0 df-count)
    (div n-docs)
    (Math/log)
    (* tf-count)))

(defn TF-IDF [src]
  (let [n-doc (first (flatten (??- (D src))))]
    (<- [?doc-id ?tf-idf ?tf-word]
      ((TF src) ?doc-id ?tf-word ?tf-count)
      ((DF src) ?tf-word ?df-count)
      (tf-idf-formula ?tf-count ?df-count n-doc => ?tf-idf))))

(defn -main [in out stop tfidf & args]
  (let [rain (hfs-delimited in :skip-header? true)
        stop (expand-stop-tuple (hfs-delimited stop :skip-header? true))
        src (etl-docs-gen rain stop)]
    (?- (hfs-delimited tfidf)
      (TF-IDF src))
    (?- (hfs-delimited out)
      (word-count src))))
```

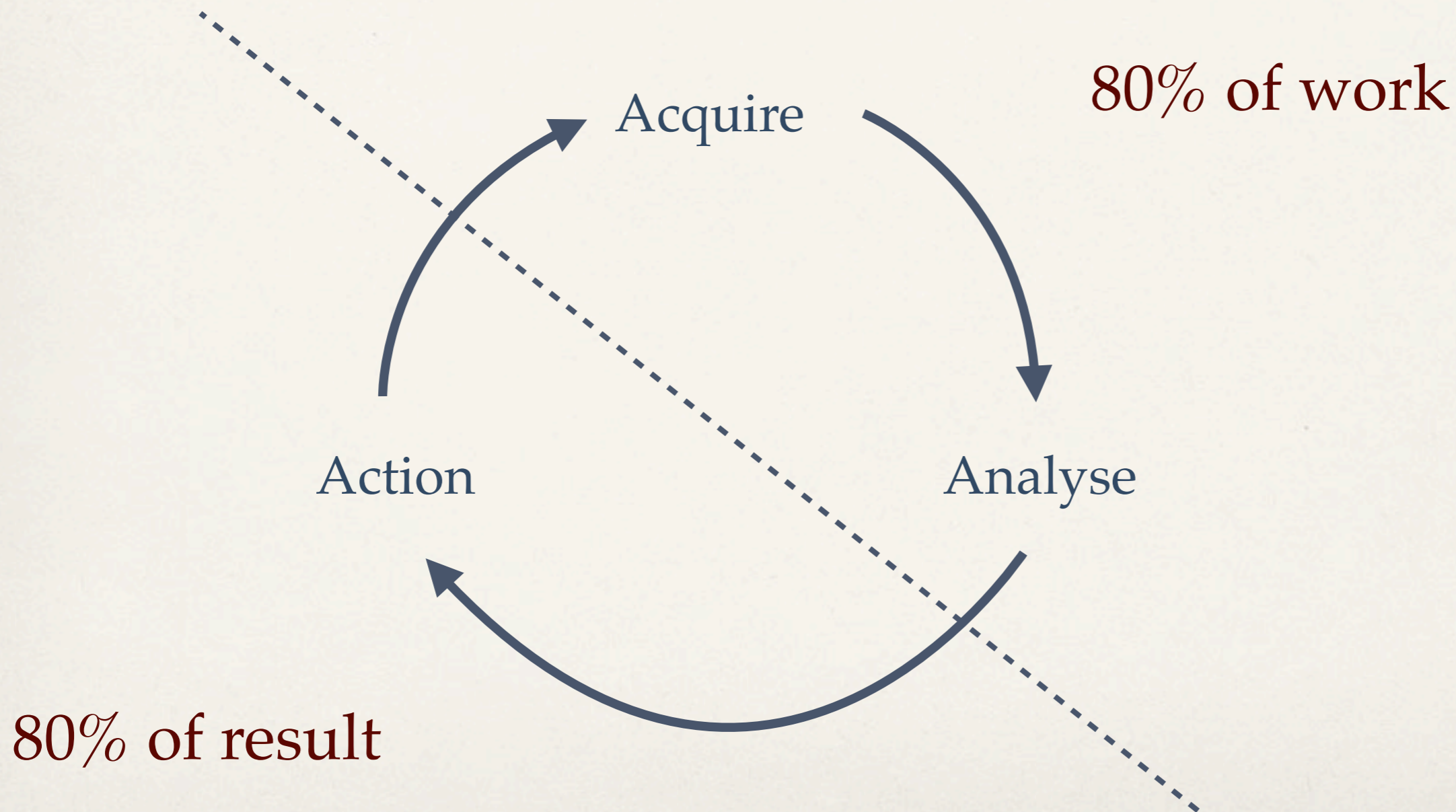

Our data processing methodology

- ❖ No monolithic framework
- ❖ Only build what we need as we go
- ❖ Composability, extensibility, maintainability



Action

80/20



Three Es

- ❖ Enlighten
 - ❖ R with rhdfs and ggplot, Sinatra + D3.js
- ❖ Expose
 - ❖ Scheduled Hadoop jobs to load processed data into MySQL for everyone to use
- ❖ Exploit
 - ❖ Real-time customer intelligence to personalise website for each visitor

Result

- ❖ Data from all levels are accessible
- ❖ Information is easy
 - ❖ "Sweet! I don't have to do anything!" -- Hemal, uSwitch developer
- ❖ Opening dialogue about using data

Summary

- ❖ Develop incrementally and iterate
- ❖ Mitigate unnecessary complexity

Contact



- ❖ Paul Lam, data scientist at uSwitch
- ❖ @Quantisan
- ❖ paul.lam@forward.co.uk