Data Engineering interview case

A glimpse into the world of Big Data at OLX

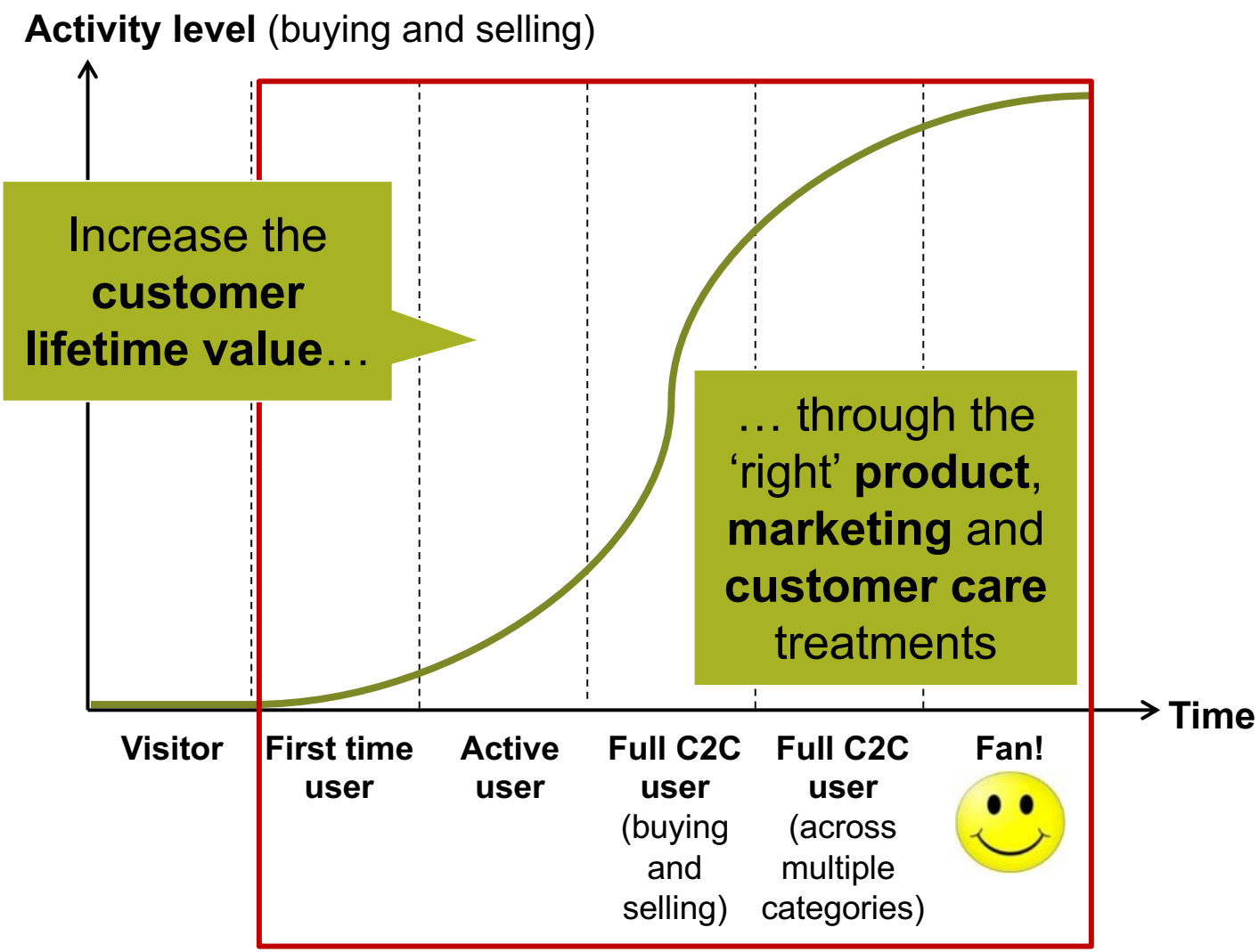
Note: Throughout this presentation we use the term **NNL = 'net new listings'**

Net = Approved, not rejected, not spam
New = First version of listing (not renewed)

Version 15

What is **C**ustomer **L**ifecycle **M**anagement?

CLM is all about **treating our customers the best way possible** in the lifecycle stage they are in, turning them into 'fans'



CLM enables:

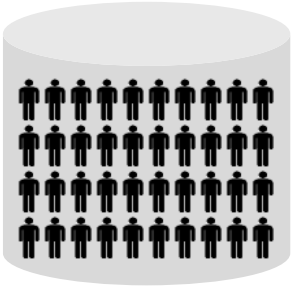
- ✓ Dramatic improvement in marketing effectiveness
- ✓ Improved retention
- ✓ Higher engagement
- ✓ Happier customers
- ✓ $xB value in the endgame

Sweeping overview of the CLM machine

1 Platforms and data

- **Raw platform data** sources
- **Multiple, disjointed data structures** – transactional, product / device, customer care, third party / social, other

2 Single customer view

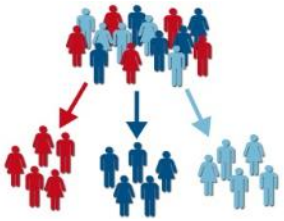


- Consolidated **CLM data warehouse** with single customer view
- **Feeds all CLM business logic** & analytics
- **Refreshed** daily but with real-time overlay

3 Insights & analytics

- **Identify where value is** and what its drivers are
- **Understand user journeys** and pain points – for buyers and sellers, by category, channel, region

4 Customer segmentation



- Classify customers according to their **lifecycle stage**
- Assign to **actionable segments** based on customer behaviour and lifetime value

5 Customer treatments

- ✓At the right **time**
- ✓For the appropriate **segment**
- ✓With the right (personalised) **message**
- ✓Via the right communication **channel** (email, SMS, push, phone, product)

6 Execution

- **Execute customer treatments**
- **Measure and analyse impact** and continuously refine CLM platform & business rules as needed

Overview of the **customer lifecycle**, from *'cradle to grave'*

Aware	1 Curious	2 Exploring	3 Trial	4 Novice	5 Repeat	6 Passive	7 Dormant	8 Lost
Aware of the OLX brand but has not yet used the product	Started using product but not yet engaged with content	Browsed product categories, searched, and / or viewed item pages	Posted / replied for the first time	Recently posted / replied for the first time	Returning customer posting and / or replying across categories	Stopped posting and replying, but still engaged with product	Stopped using product for a period of time	"Perma-nently" stopped using product

Increase engagement

Reduce churn

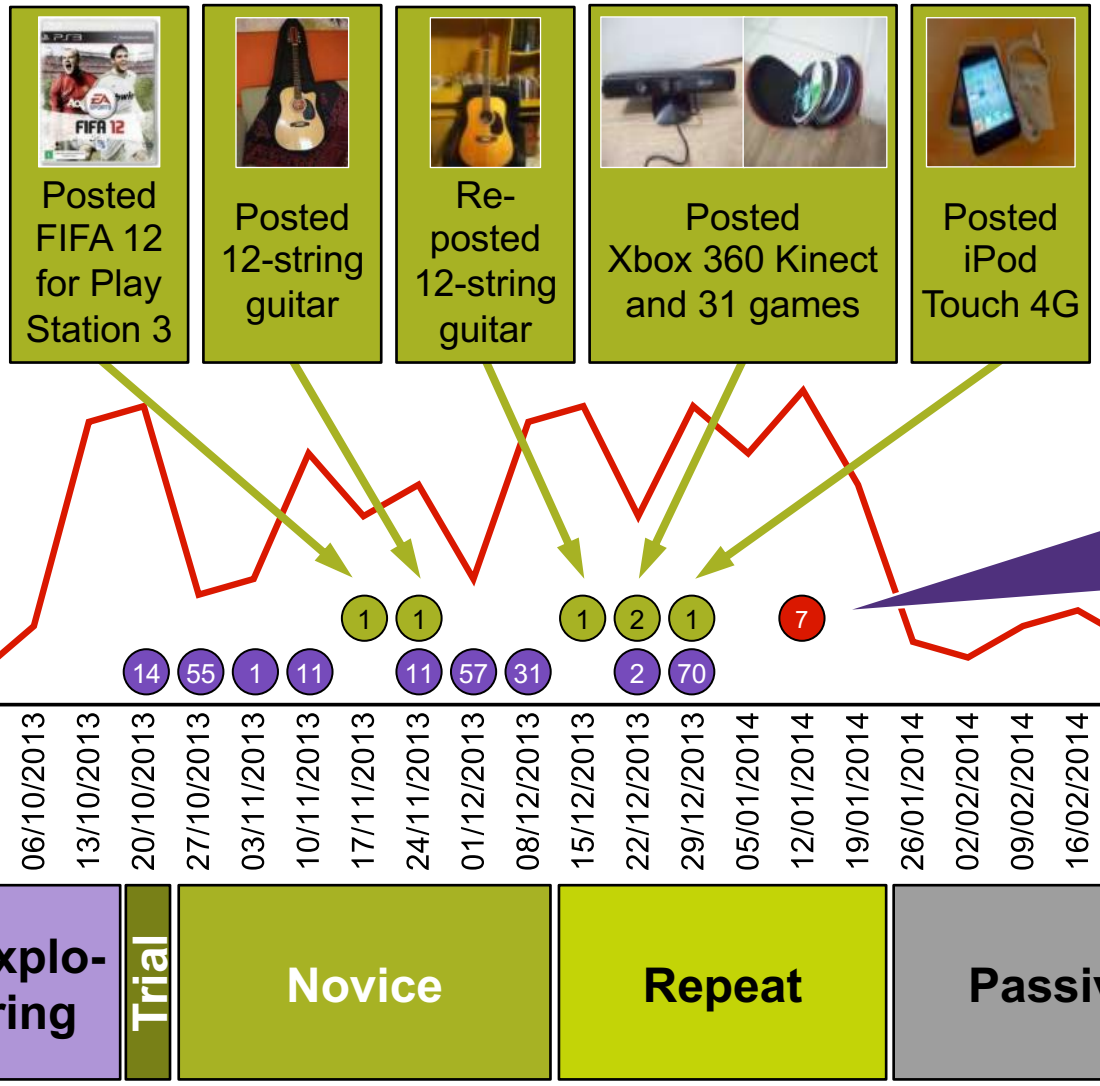
Customer needs and business objectives evolve over the lifecycle

	Aware	1 Curious	2 Exploring	3 Trial	4 Novice	5 Repeat	6 Passive	7 Dormant	8 Lost
Customer needs	*"**I am not sure** online classifieds is for me."*	*"I am **looking for a good** online classifieds platform to buy / sell something."*	*"I want to **find** what I am looking to buy."* *"I want to ensure I would be able to **sell my item**."*	*"I want an **easy, effective and safe** experience."*	*"I completed my first sale / purchase... **now what?**"* *"I **need reasons** to use the platform again."*	*"I **need ideas and tips** for what I can buy / sell."* *"I **love** buying and selling stuff with OLX."*	*"I **need help** finding what I am looking for."* *"I need **support** with the posting process."*	*"I need **better reasons or incentives** to keep me interested in using the platform."*	*"I **don't want to use** an online classifieds platform."* *"I **don't like** the OLX experience."*
Business objectives	**Attract**	**Impress**	**Capture**	**Delight**	**Activate**	**Grow**	**Re-activate**	**Revive**	**Recover**
	…through **online and offline marketing** acquisition channels	…through providing **superior product experience** and propo-sition	…via persona-lised and highly relevant **product journeys**	…with **intimate support** and hand-holding during trial process	…by **nurturing relation-ship** and driving towards repeat usage	…engage-ment through **cross/up-selling and loyalty marketing**	…with tailored **proactive retention campaigns**	…with **targeted reactive retention**	…**opportu-nistically** (but do not target explicitly)

Illustration of a (real) customer lifecycle

ID: 16342062; email: djguto00@gmail.com

augusto cesar simplicio carlos
@djguto00
Sao paulo
http://Djguto

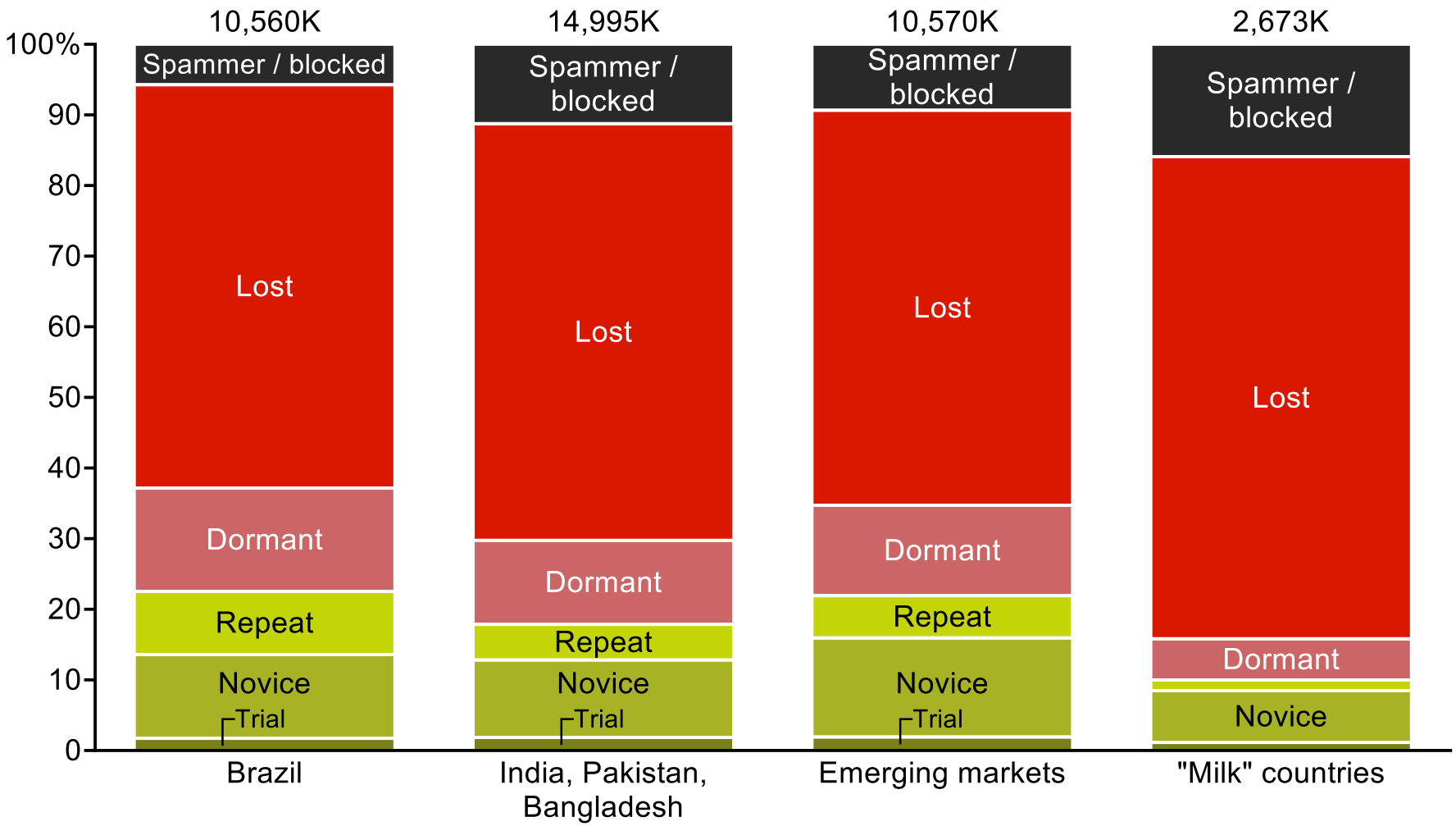


A technology buff, a music lover and a gamer, Augusto was one of our most active Repeat customers, before being blocked for contacting too many sellers in 1 day

Source: OLX BI data warehouse
Note: Visits is dummy data

Today's snapshot: **65-70%** of our customers are inactive (**Dormant** + **Lost**) and only **5-10%** are coming back regularly (**Repeat**)

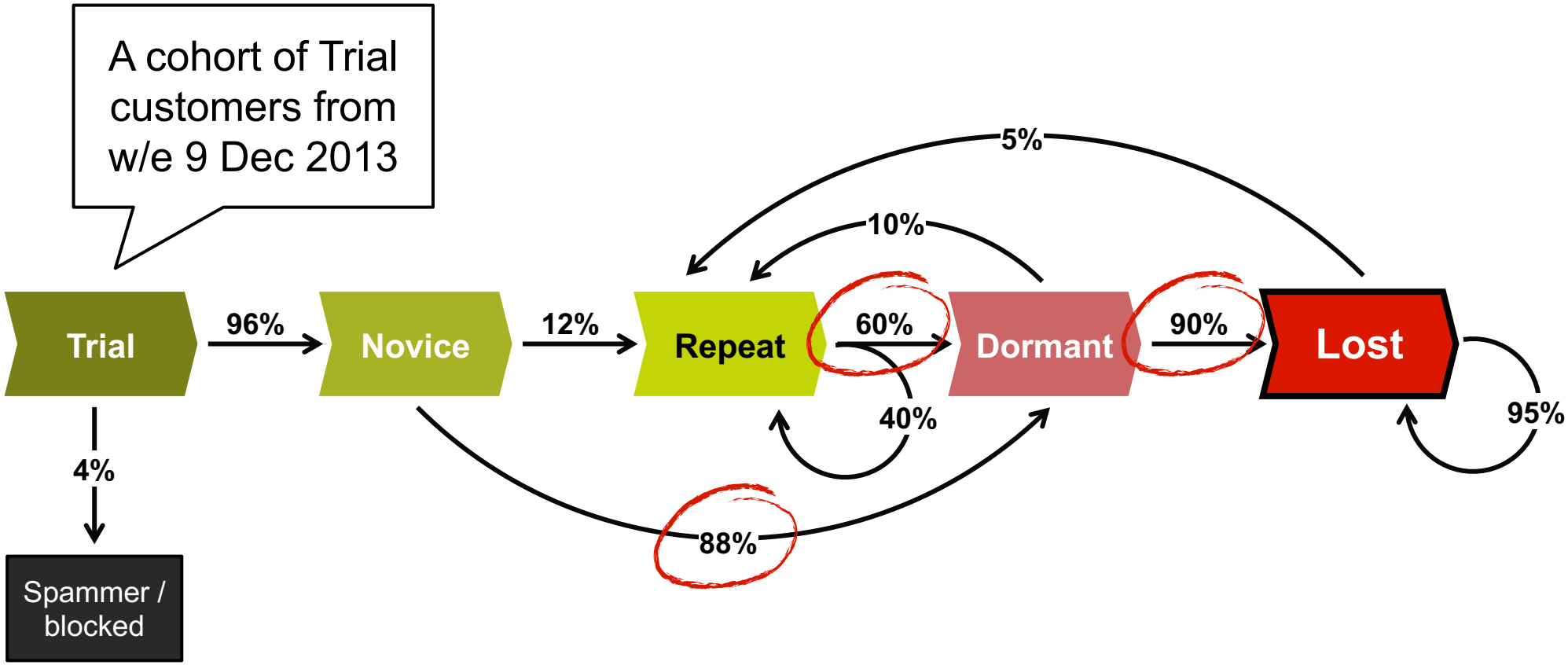
customers by region (top 24 countries), as of 29 June 2014



There is significant customer 'leakage' across all lifecycle stages

Brazil

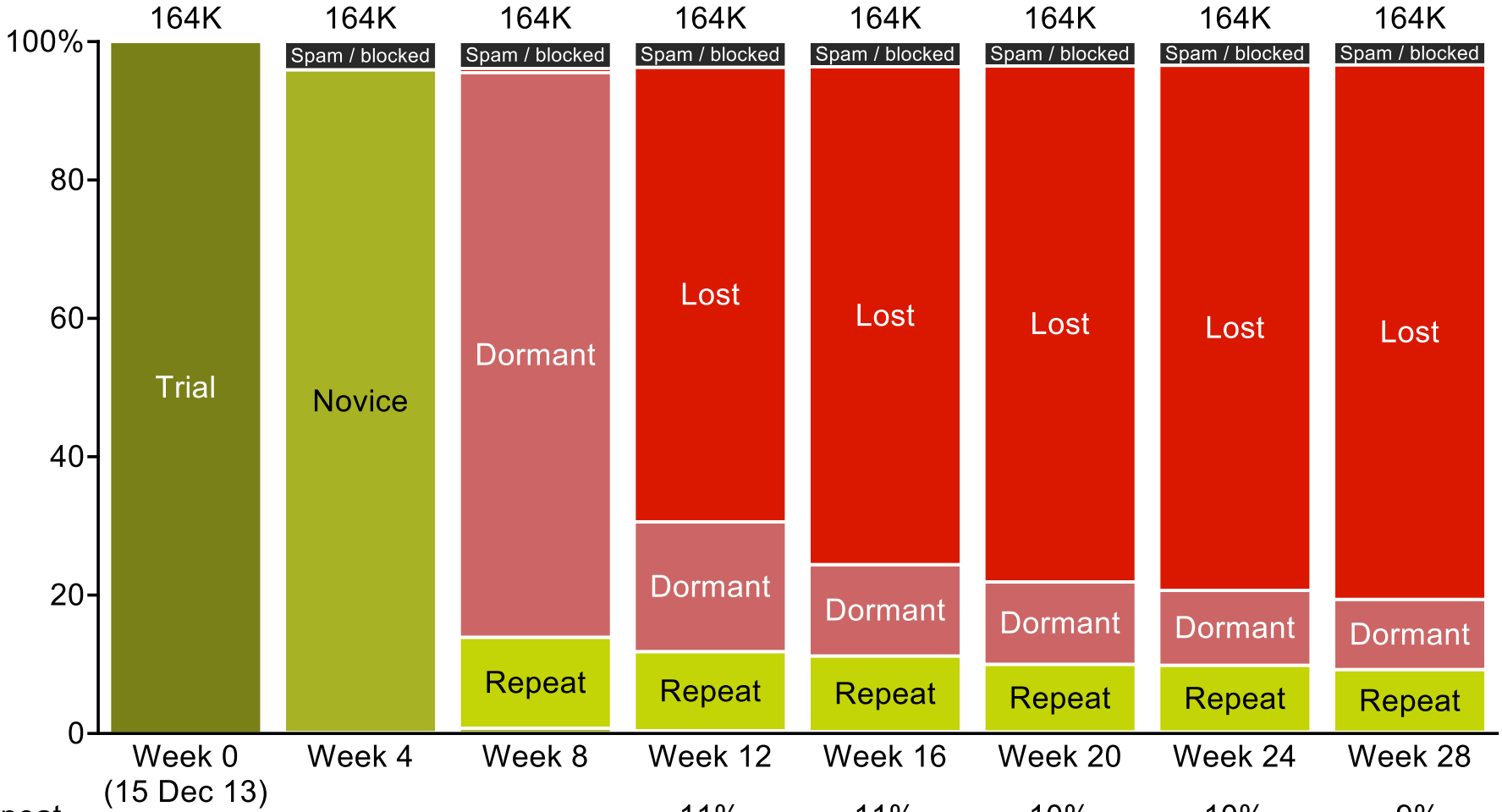
Behavioural propensities by lifecycle stage



Source: Internal data

Impact over time: only **~9% of Trial** customers remain active after ~6 months, **77% end up Lost**

Evolution of Trial customers, cohort w/e 15 Dec 2013, 000's

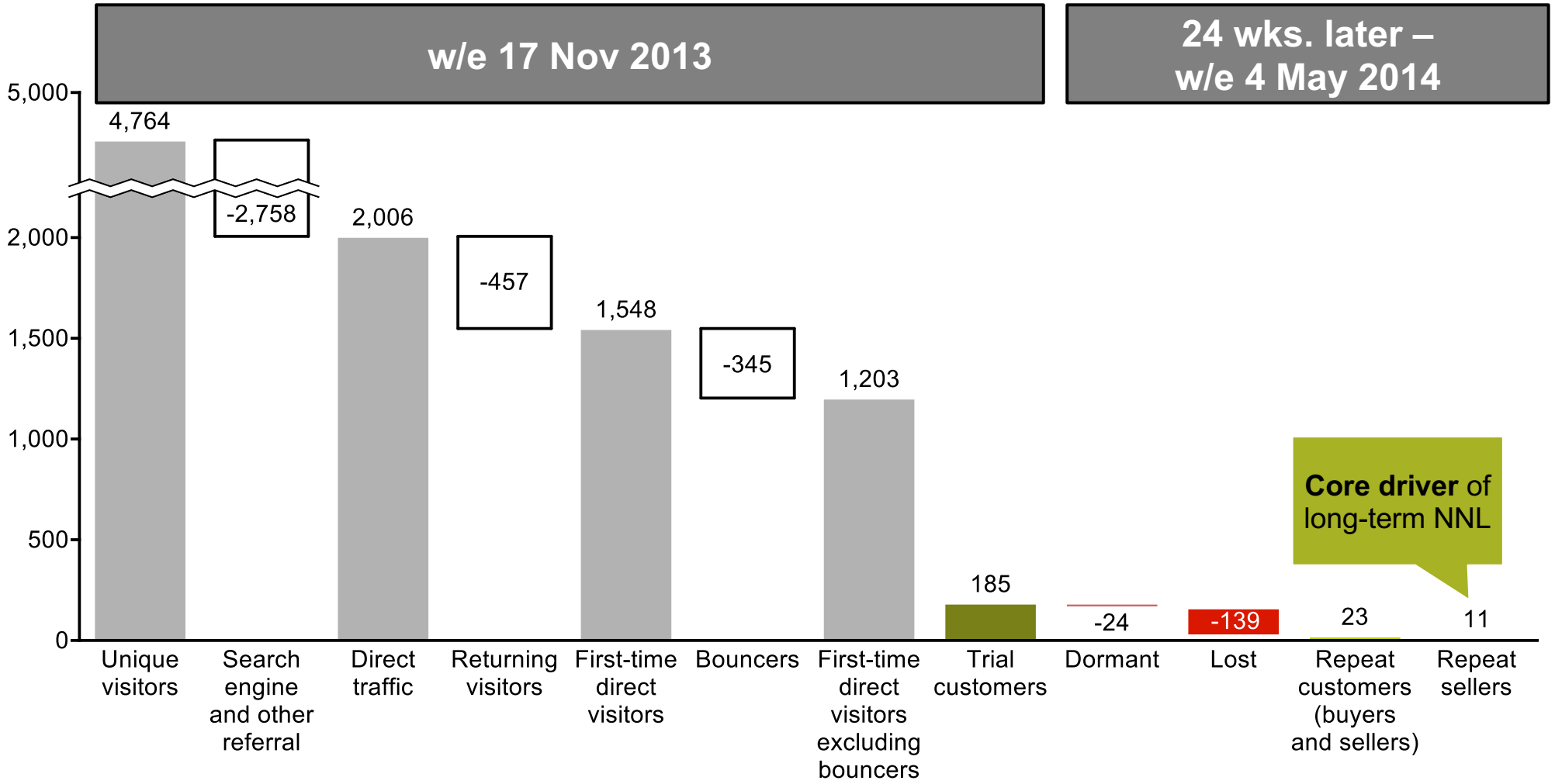


	Week 12	Week 16	Week 20	Week 24	Week 28
% Repeat	11%	11%	10%	10%	9%
% Lost	66%	72%	75%	76%	77%

Source: OLX BI data warehouse
Note: Based on extrapolated 10% Brazil sample. *Active = New + Trial + Repeat

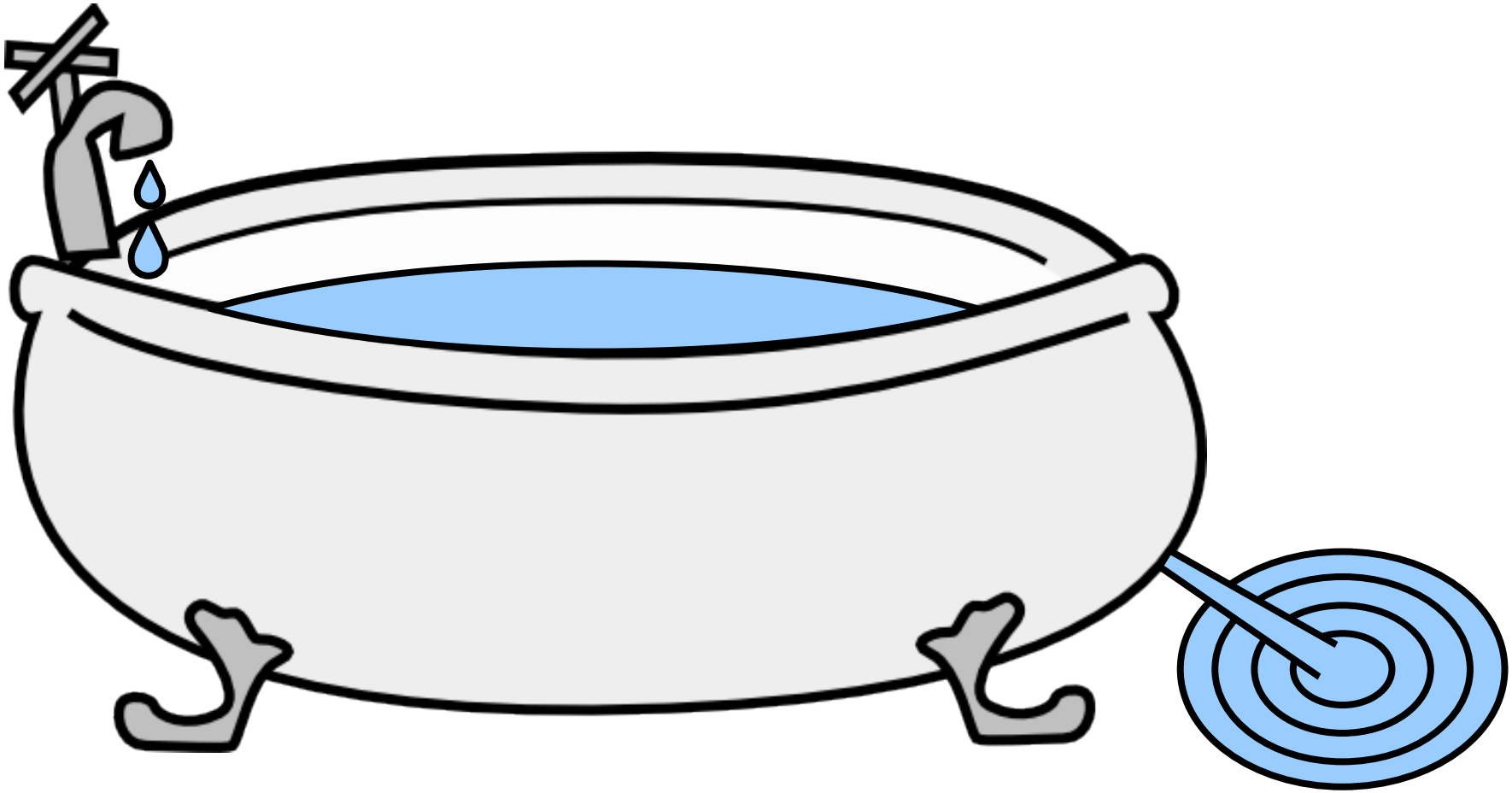
Ultimately very **few customers become Repeat sellers** – huge opportunity to improve conversion across the lifecycle

Lifecycle funnel, # customers, 000's (excludes mobile app usage)



Source: AT Internet; OLX BI data warehouse; internal marketing spend data; analysis
Note: Usage data based on extrapolated 10% Brazil sample and excludes usage from mobile app users. Marketing spend in Nov 2013 in Brazil was $6.3M.

So… how do we fix the leakage? Let's look at 5 opportunities…



5 strategic CLM objectives to fix this, leveraging (email, push and phone) marketing, product changes and customer care to get there

Objective	Current state	Target
1 Capture Exploring customers	• Only **15%** of Exploring customers become new Trial users (to be validated)	**17.5%**
2 Activate Novice customers	• Only **12%** of Novice customers become active Repeat users	**17%**
3 Retain Repeat customers	• **40%** of Repeat customers remain active after 6 months	**50%**
4 Grow Repeat customers	• **47%** of Repeat customers are sellers (53% are buyers only) • **3 NNL** on average per active Repeat seller per month	**50%** **3.3 NNL**
5 Recover Dormant / Lost customers	• **10%** of Dormant customers (5% of Lost) re-activate and become Repeat customers again over 6 months	**20%**

Impact

Double daily NNL rate

Fun exercises for potential new Data Engineers

Objectives	• Load data to Redshift • Implement CLM segmentation on sample customer data • Calculate listing liquidity
Deliverable	• Table in Redshift cluster • SQL code
Format	• Redshift SQL

Some tips

- Read the Redshift documentation ☺
  http://aws.amazon.com/documentation/redshift/
- Recommended SQL clients:
  - **Datagrip**
    https://www.jetbrains.com/datagrip/
- The given data is only a small sample of the real dataset (which is 500x bigger), so think carefully about performance of your query as it is supposed to still run fast on a bigger dataset.
- Depending on the size of your test cluster, your queries should usually finish in less than a couple of minutes minutes

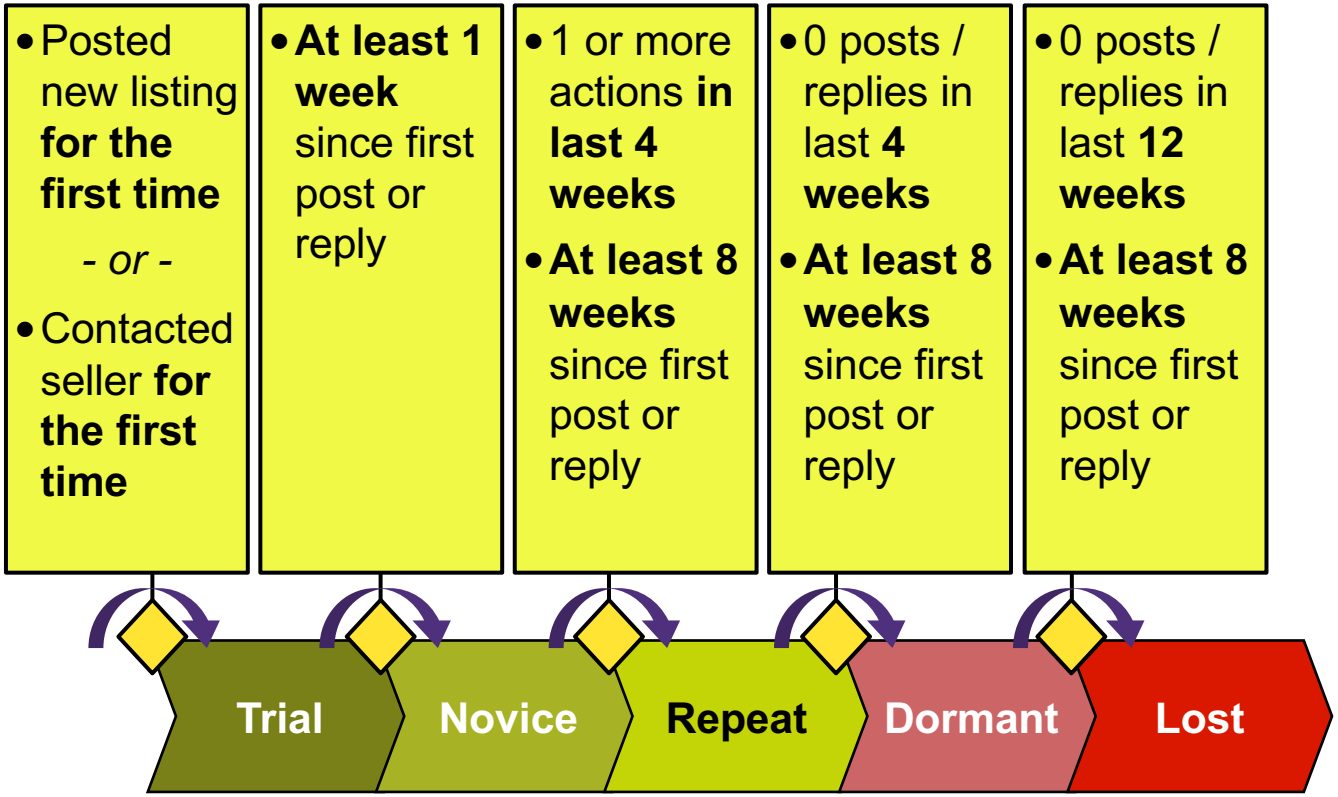
Description of data structure of sample table **actions**

Name	Type	Values
user_id	**VARCHAR(48)**	Unique id of each user that executed the action
action_type	**CHAR(1)**	One of two possible actions: ‘P’ – post (create new ad); ‘R’ – reply (contact seller)
action_ts	**TIMESTAMP**	Date-time timestamp of the action
item_id	**VARCHAR(48)**	The unique id of the item that was posted or replied to
device	**VARCHAR(2)**	Channel / device used for action: ‘DW’ – desktop web; ‘MW’ – mobile web; ‘MA’ – mobile app; ‘U’ – Unknown
b2c	**BOOLEAN**	Whether the item belongs to a B2C product category (Jobs, Real Estate, Vehicles) 1 – item is in B2C category NULL – item is not in B2C category

Exercise 1: Load sample data from S3 to a new Redshift cluster

- Read Redshift documentation
- Set up free Redshift cluster
- Load sample data (stored in publicly accessible AWS S3 directory)
  - Bucket=s3://tradus-bi-recruiting/actions/
  - Size=5 GB compressed (gzip)
  - aws_access_key_id=AKIAIQGLDQCPBKH6VQCA
  - aws_secret_access_key=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
  - region: eu-west-1
  - You can list the files using AWS CLI using:
    `aws s3 ls s3://tradus-bi-recruiting/actions/`
  - Files were unloaded from Redshift using:
    `UNLOAD ... PARALLEL ON GZIP`
- Tip: Think about distribution, sort keys and encoding!

Exercise 2: Segment customer base for below 5 lifecycle stages



Note: **action** = post -or- reply

Outputs:

- A table with the designated segmentation for each user as of 1 July 2018 + SQL code
  TIP: *Use daily precision (not weekly precision)*
- A query that shows the distribution (relative size) of each lifecycle stage based on this table

Exercise 3: Calculate liquidity

Definitions:

- **Replies within X days** → Number of replies received within X days of an item's posting date
  - *Example: **'Replies within 7 days'** is number of replies received (calculated for each item) between the item's posting date and (posting date + 7 days)*
- **Liquid items X replies within Y days** → Number of items posted on a given date that receive X or more replies within Y days
  - *Example: **'Liquid items 3 replies within 7 days'** for a given date is the number of items with 3 or more replies received within 7 days since posting date*
- **% liquidity X replies within Y days** → % of all items posted on a given date that receive X or more replies within Y days
  - *Example: **'% liquidity 3 replies within 7 days'** for a given date is the number of items with 3 or more replies received within 7 days since posting date ÷ All ads posted on that date*

Exercise:

- Write a query that creates a fact table **fact_item_liquidity** with the following information:
  - **Dimensions:**
    - **date** (all dates available in dataset)
    - **item_id** (all items available in dataset)
  - **Measures:**
    - # replies received within 1 day
    - # replies received within 7 days
- Write a query that creates a fact table **fact_liquidity** with the following information (Tip: Use previous table as input):
  - **Dimensions:**
    - **date** (all dates available in dataset)
  - **Measures:**
    - # items posted on date
    - Liquid items 1 reply within 1 day
    - Liquid items 3 replies within 7 days
    - Liquid items 5 replies within 7 days
    - % liquidity 1 reply within 1 day
    - % liquidity 3 replies within 7 days
    - % liquidity 5 replies within 7 days

Final notes

- Sample table is highly simplified for this case. As simple as it might look, it will give us a good insight in your code style and coding skills.
- Comment your code and make it readable . Use 4 spaces or tabs that represent 4 spaces. See sample formatting below.
- Send your output in a zip file including access details to your test cluster and console
- If in doubt make sure to clarify exercises with your interviewer to avoid risk of incorrect output

```
-- Example code formatting
SELECT  x,
        y,
        MAX(z)
  FROM  a
  JOIN  t ON a.x = t.x
  LEFT  JOIN t2 ON a.x = t2.x
 GROUP  BY x,y
HAVING  MAX(z) > 1
 ORDER  BY x,y
 LIMIT  100;
```