Hello!

I am Alex Egorenkov

Can we predict who will retweet #personalvictory

1

MOTIVATION

INITIAL IDEAS

I wanted to work with…

- Text Data
- Stream Data
- Anything but policy

GA COMMUNITY

I ran into...

- John
- www.personalvictories.com
- Displays #personalvictory tweets

> *Learning to use git. No longer afraid to open Terminal.* **#personalvictory**

John does not understand his viewers fully.

What can data science say about his viewers?

START SMALL

Can we predict...

- who will post personal victories?
- who will favorite personal victories?
- who will retweet personal victories?
- how many statuses will be posted?

2

INITIAL WORK

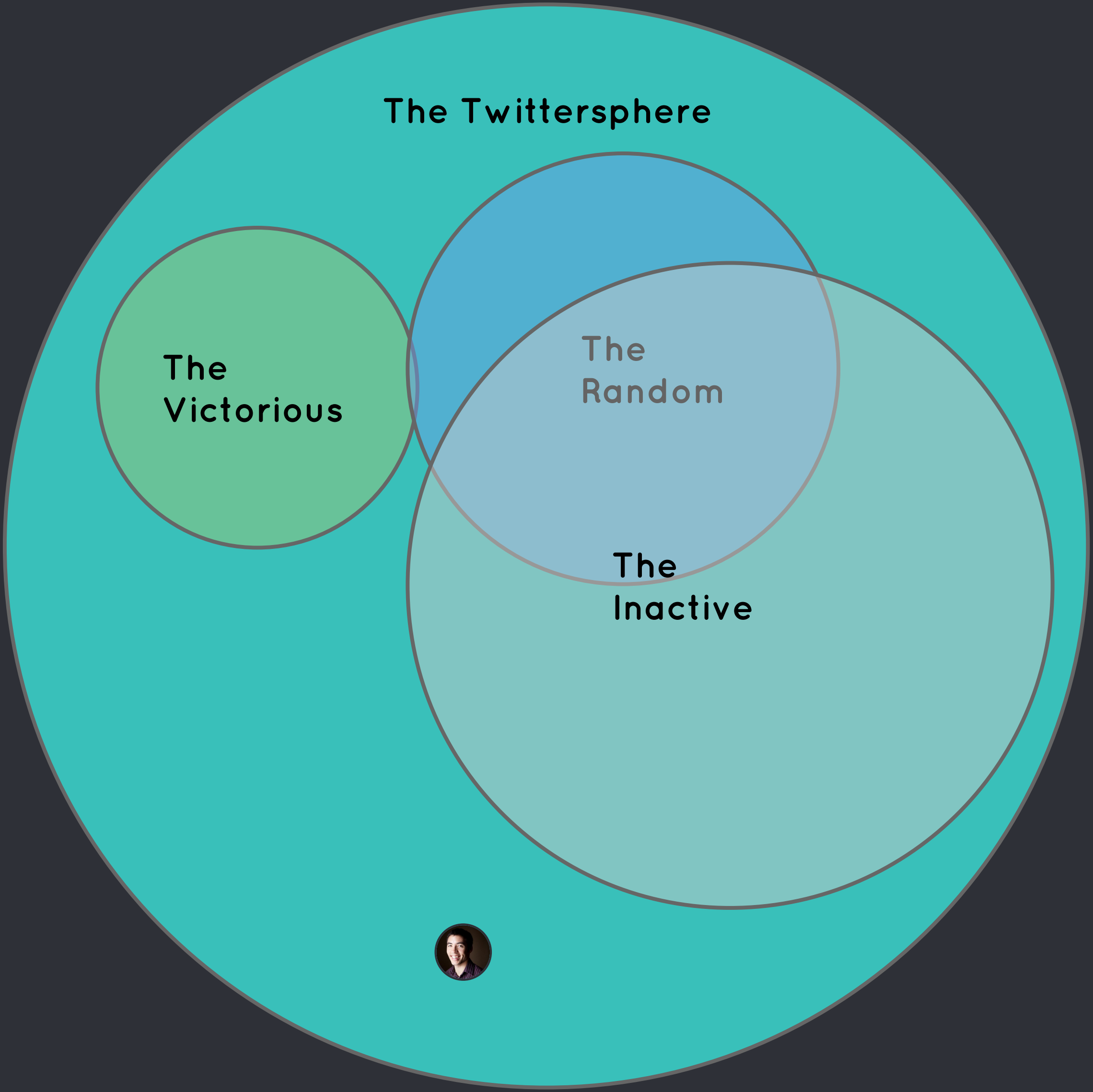
THE EASIEST QUESTION

- Who will post personal victories?

THE APPROACH

user1	followers count	tweet keyword1	tweet keyword1
user2	followers count	tweet keyword1	tweet keyword2
		⋮	
user2000	followers count	tweet keyword1	tweet keyword1
user2001	followers count	tweet keyword1	tweet keyword2

THE APPROACH



THE APPROACH

✓ Get random sample of users who post personal victories

✓ Get random sample of users who don’t

✓ Explore

❏ Classification model with simple count features

❏ Text based classification model

❏ Validate with ROC/AUC due to class imbalance

❏ Repeat with larger question

THE DATA

Currently …

- Account creation date
- User description
- Favorites Count
- Followers Count
- Friends Count
- Language
- List subscriptions
- Location
- Last status posted by user

Will collect …

- Last 3,000 statuses posted by user

3 CHALLENGES

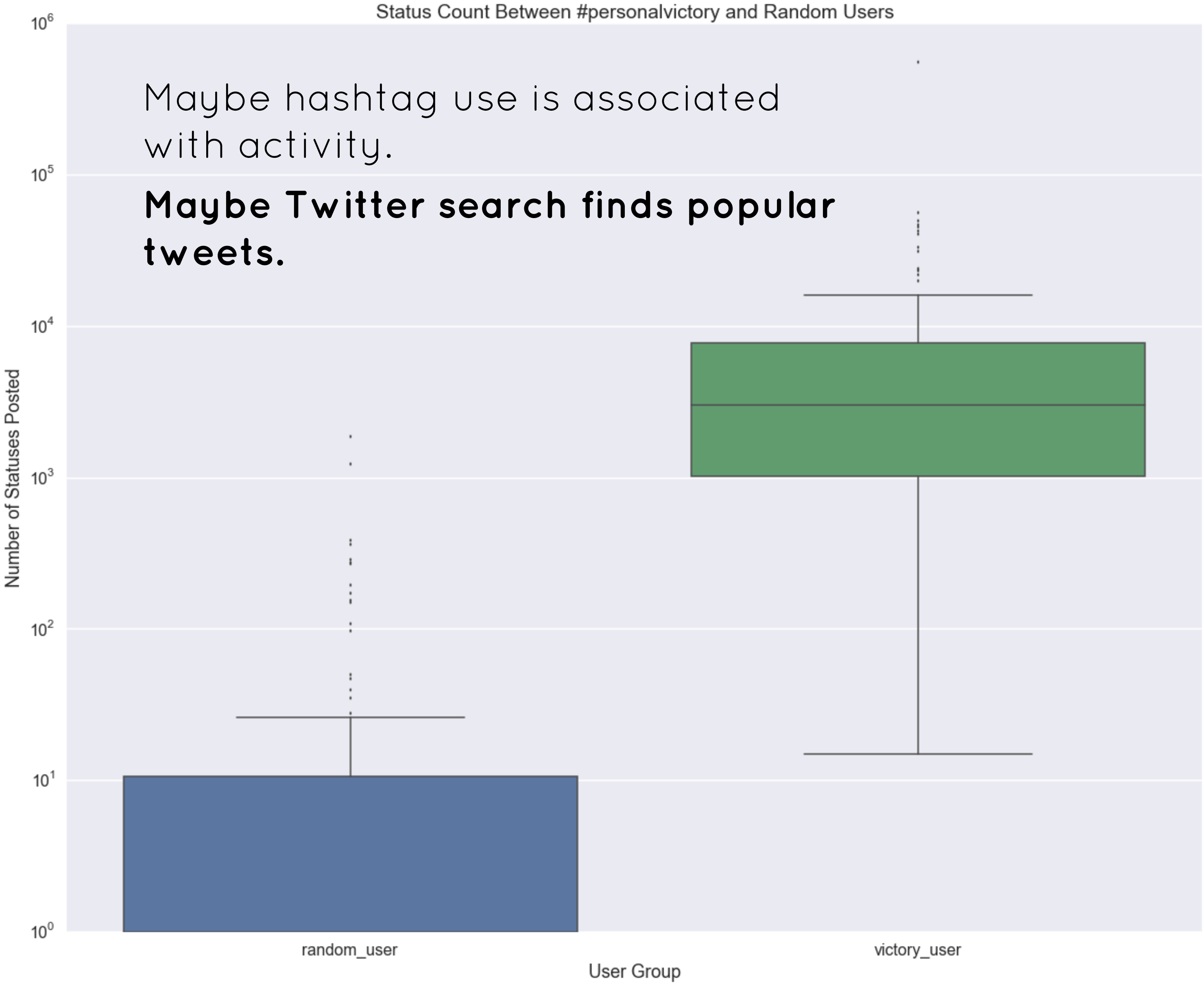
THE TWITTER API IS FRUSTRATING

- Search index only goes back to 6-9 days
- Rate Limits:
  - user lookup: 180 request/15 minutes
  - retweet lookup 15 requests/15 minutes
- No GET request for favorites

It will take 48 hours to retrieve retweeting users.

(Let's call it 96 hours to account for mess-ups)

- Unicode/codec errors
- Non-random missing data
  - Many users in the random group have not posted a status
- Sample issues
  - 3000 victories may be a small sample for text data
  - Twitter search is not so random



Maybe hashtag use is associated with activity.

Maybe Twitter search finds popular tweets.

> *Eighteen slides in five minutes.*
>
> **#personalvictory**

Thanks!

SUGGESTION ARE WELCOME!

CREDITS

Special thanks to all the people who made and released these awesome resources for free:

- Presentation template by SlidesCarnival
- Photographs by Unsplash