

# CSE 190 – Lecture 2

Data Mining and Predictive Analytics

Using Regression to Predict Content

Popularity on Reddit

# Images on the web

To predict whether an image will become **popular**, it helps to know

- Its **audience**, or the **community** it was submitted to
- Whether it is **original** compared to previous content
- How it was **marketed** (e.g. its posting title)

(e.g. Bandari et al. 2012; Artzi et al. 2012; Hogg & Lerman 2010; Lee et al. 2010; Petrovic et al. 2011; Tatar et al. 2011, and others)

# Predicting success of content on image-sharing communities



This photo recently one the Andrews award for the 'most perfect timing of a Nature photograph', I can see why.

submitted 29 days ago by SICK\_OF\_ to /r/pics

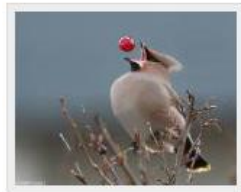
11 points  
1 comment



NOM! (Photo by: Bohemian Waxwing)

submitted 2 months ago by favoritehell [deleted] to /r/PerfectTiming

1117 points  
1 comment



Perfect moment bird (ex-post from r/pics)

submitted 25 days ago by 123imAwesome to /r/photoshopbattles

36 points  
1 comments



A bohemian waxwing eating a berry

submitted 4 months ago by HazeySynth to /r/pics

39 points  
1 comment



Bird shot at the perfect moment

submitted 25 days ago by arbili to /r/pics

2712 points  
166 comments



Perfect timing.

submitted 4 months ago by animalpath to /r/pics

2555 points  
71 comments



Perfect timing.

submitted 2 months ago by presaging to /r/aww

12 points  
1 comment



Timing is Everything

submitted 5 months ago by Xnicko378X to /r/pics

10 points  
1 comment

**“Who will like my content, and how should I market it?”**

# Resubmissions on reddit.com

When social media content is posted,  
can we determine

How much of the  
success was due to  
the **content itself**

vs.

How much of the  
success was due to  
how the content  
was **marketed**

Why?

Changing how content is **presented** is easier than  
changing the content itself!

# Resubmissions on reddit.com

↑ [I'm not sure I quite understand this piece](#)

62 Submitted 2 years ago to pics by xxx

↓ 24 comments

↑ [How wars are won](#)

20 Submitted 18 months ago to WTF by xxx

↓ 1 comment

↑ [Murica!](#)

774 Submitted 1 year ago to funny by xxx

↓ 59 comments

↑ [Bring it on England, Bring it on !!](#)

10 Submitted 10 months ago to pics by xxx

↓ 4 comments

↑ [I believe this is quite relevant currently](#)

226 Submitted 7 months ago to funny by xxx

↓ 15 comments

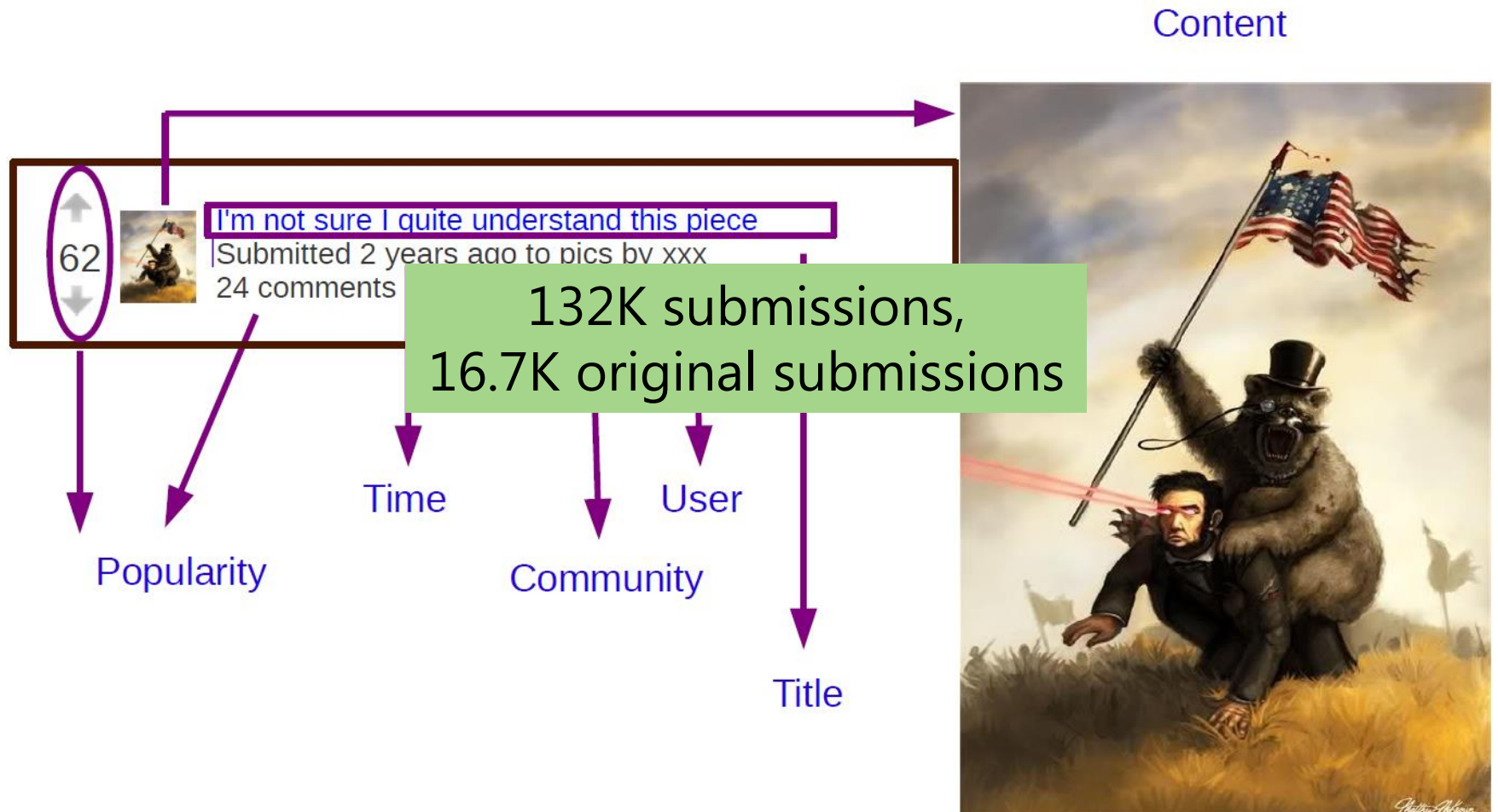
↑ [God bless whoever makes these](#)

794 Submitted 1 month ago to funny by xxx

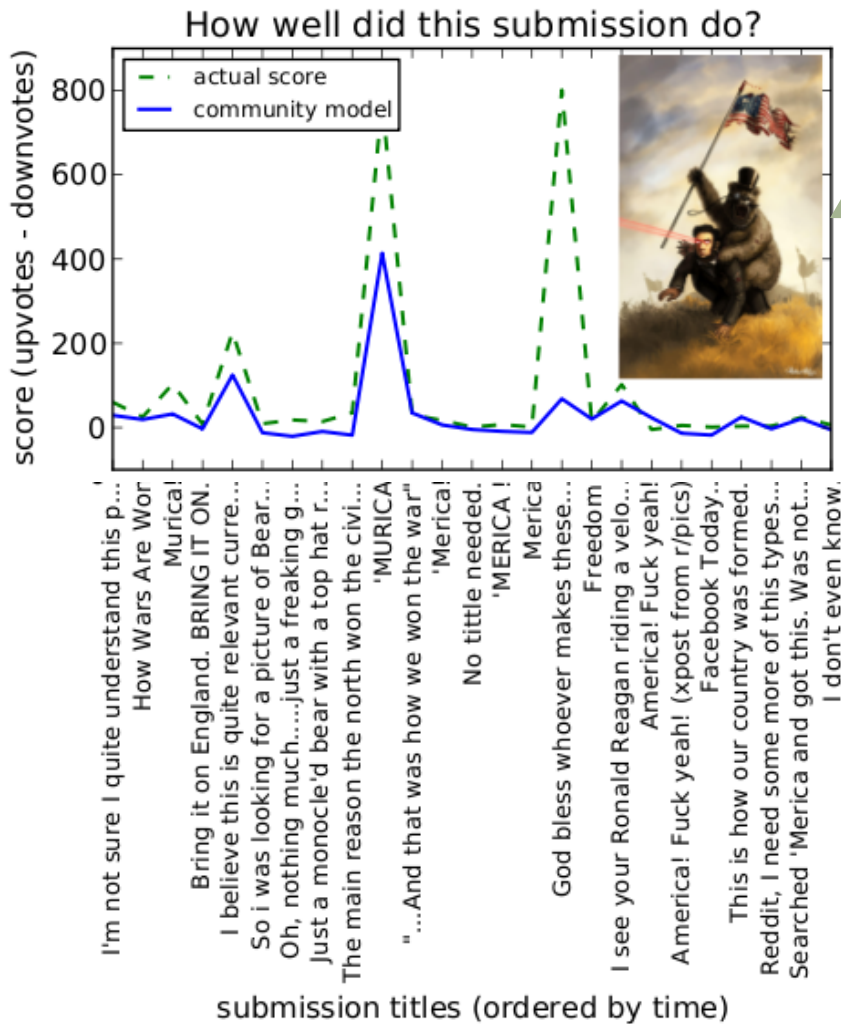
↓ 34 comments



# Understanding popularity

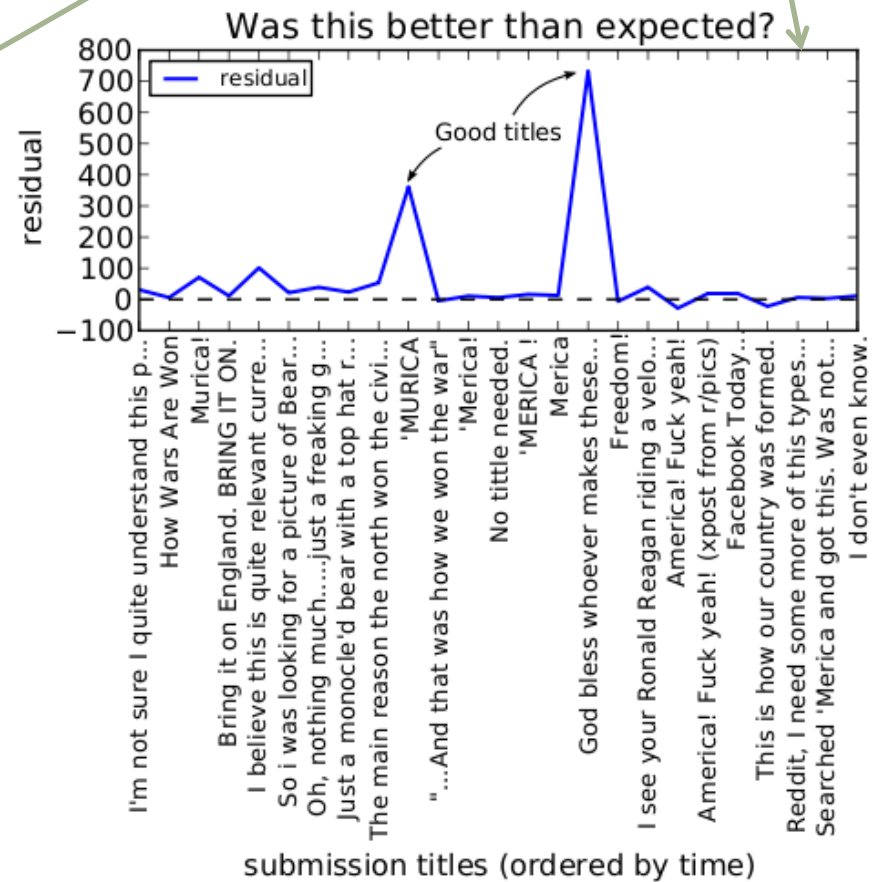


# Resubmissions on reddit.com

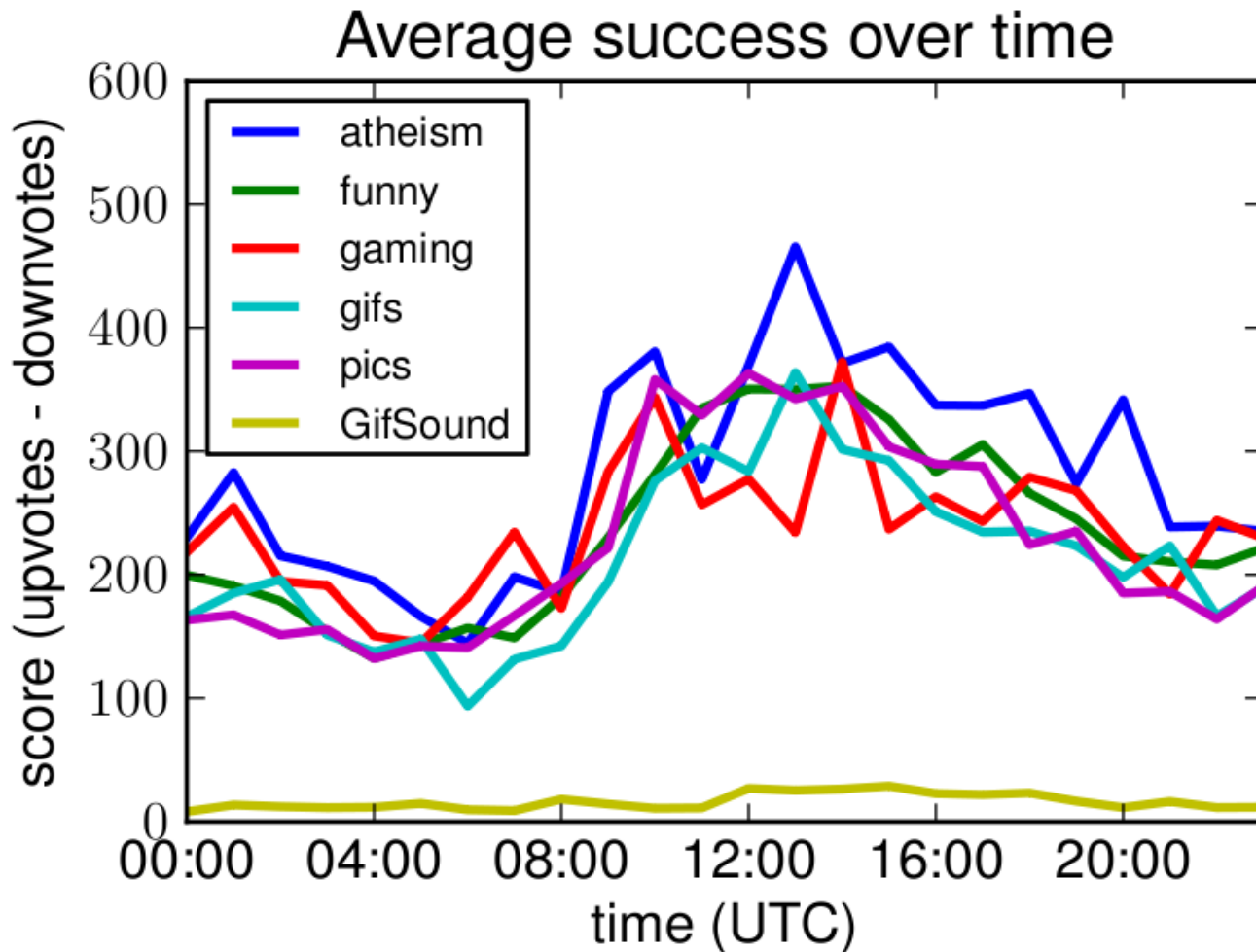


## Community effects

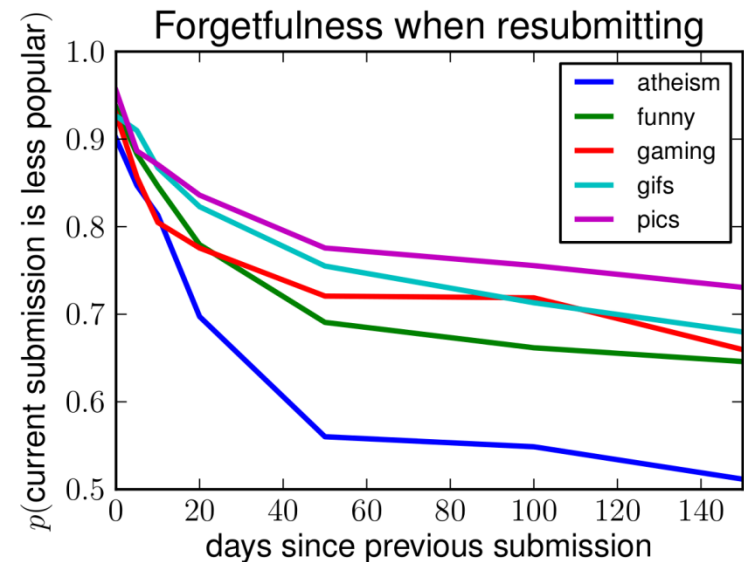
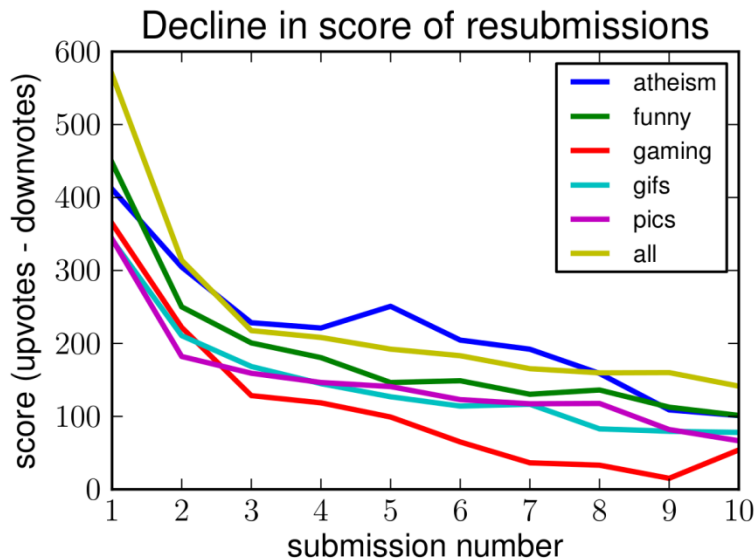
## Language effects



# Temporal effects on reddit



# Temporal effects on reddit



Resubmissions are less popular (left), but can still be popular if we wait long enough (right)



# Model (non-title effects)

$$\hat{A}_{h,n} = \underbrace{\beta_h + \phi_h}_{\text{inherent popularity}} \underbrace{\exp}_{\text{decay from resubmissions}} \left\{ \underbrace{-\sum_{i=1}^{n-1} \frac{1}{\Delta_{i,n}^h}}_{\text{forgetfulness}} \underbrace{(\delta(c_{h,i} \neq c_{h,n})\lambda_{c_{h,i}} + \delta(c_{h,i} = c_{h,n})\lambda'_{c_{h,i}})}_{\text{other communities}} \underbrace{A_{h,i}}_{\text{previous submissions}} \right\}$$

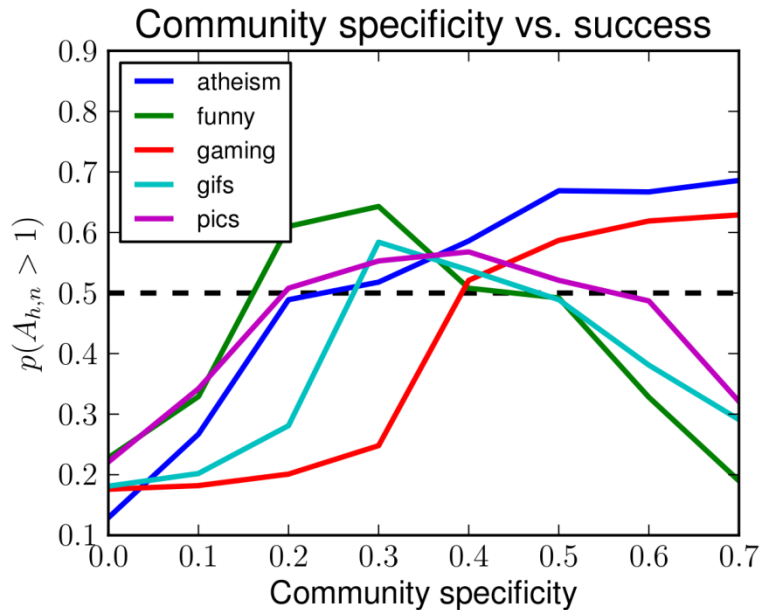
same community twice

The model is designed to account for five factors:

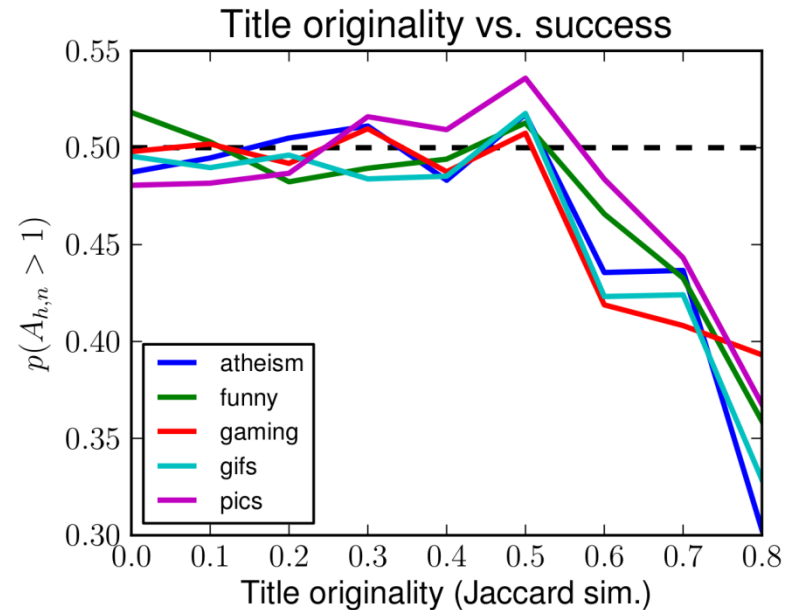
1. The inherent popularity of the content (i.e., factors other than the title)
2. The decay in popularity due to resubmitting the content
3. This decay should be discounted for old enough submissions
4. A penalty due to resubmitting to another community
5. A penalty due to resubmitting to the same community twice

(we also account for other factors, such as the time of day etc.)

# Model (title effects)



Titles should match others in the same community, but should not be too similar



Titles should differ from those previously used for the same content

# Regression, and *in situ* evaluation

Performance on held-out test data:

Model	R <sup>2</sup>
Community model only	0.528
Language model only	0.081
Community + language	0.618

- We generated **pairs of titles** for 85 submissions, which we submitted simultaneously to two different communities
- The 'good' titles garnered three times as many upvotes as the 'bad' ones (10,959 vs. 3,438)
  - Five good titles reached the front page of their community, and two reached the front page of r/all

# Example



- **Good title: What I would do to someone I hate**

- Votes: 7087+ 5228-, Cmts: 518

- **Why is this good?**

- Original title
- Optimal length (not too short)
- POS tags: Interesting (uncommon) sentence structure compared to a flat-tone syntax

- **Bad title: Funny gif**

- Votes: 300+ 124-, Cmts: 9

- **Why is this bad?**

- Not original, too generic (no specificity)
- Short length
- Flat POS tag distribution

# Conclusion

- To understand whether a submission will succeed we must understand the **content** but also their **context**
  - **When** was the image uploaded?
  - To **which community** was it submitted?
  - What is its **title**?
- We showed that context can be used to predict what will “go viral” on social media
- See the paper on <http://cseweb.ucsd.edu/~jmcauley/pdfs/icwsm13.pdf>
- Joint work with Himabindu Lakkaraju and Jure Leskovec