Please

Ask questions through the app

Rate Session

Thank you!

Machine Learning in the Wild: Techniques for Understanding Your Audience

Sarah Guido
Senior Data Scientist, Mashable

GOTO Berlin 2017

Who am I?

- Senior Data Scientist at Mashable
- NYC Python co-organizer
- Conference speaker
- O’Reilly Media author
- @sarah_guido

Agenda

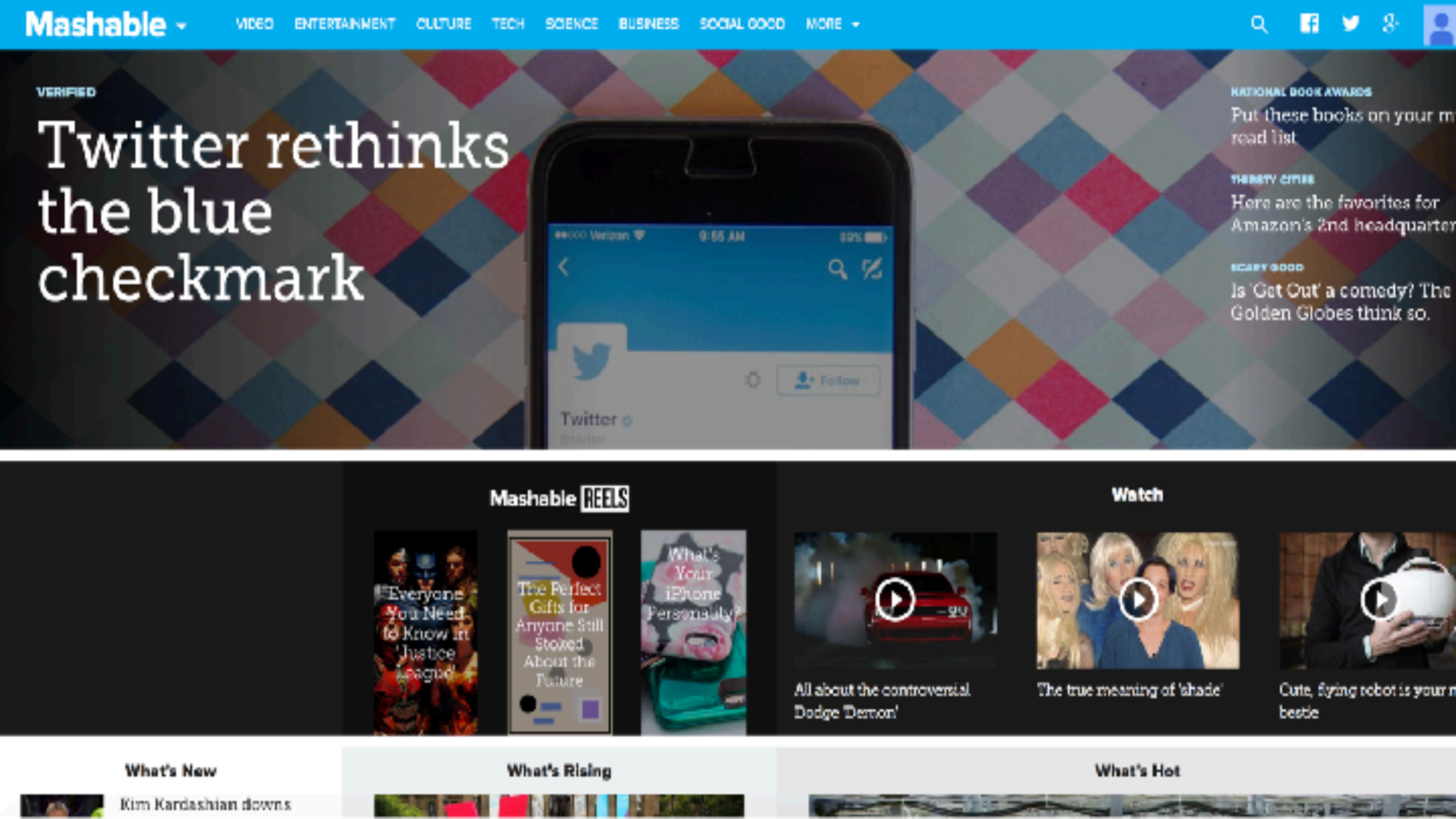
1 About Mashable

2 Content engagement

3 Audience segmentation

4 Social media strategy

5 Wrap up


Mashable
VIDEO
ENTERTAINMENT
CULTURE
TECH
SCIENCE
BUSINESS
SOCIAL GOOD
MORE
VERIFIED
Twitter rethinks the blue checkmark
NATIONAL BOOK AWARDS
Put these books on your m
read list
THIRSTY CITIES
Here are the favorites for
Amazon's 2nd headquarter
SCARY GOOD
Is 'Get Out' a comedy? The
Golden Globes think so.
Verizon
9:55 AM
Follow
Twitter
Mashable REELS
Everyone You Need to Know in 'Justice League'
The Perfect Gifts for Anyone Still Stoked About the Future
What's Your iPhone Personality?
Watch
All about the controversial Dodge 'Demon'
The true meaning of 'shade'
Cute, flying robot is your n
bestie
What's New
Kim Kardashian downs
What's Rising
What's Hot

About Mashable

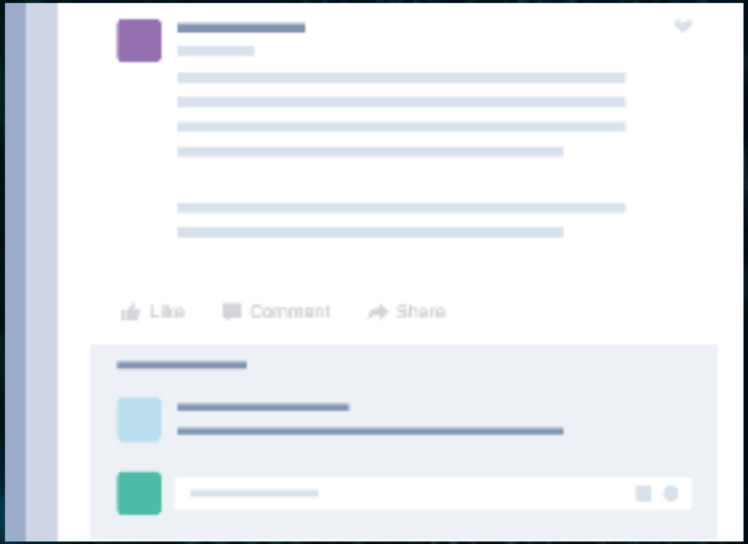
Mashable is a media and entertainment company for superfans. We’re not for the casually curious. We devour culture and tech. Our ideas shape the future and we speak to new influencers -- the early adopters who obsess with us around the globe.

Our proprietary Velocity technology suite gives us the unique ability to combine creativity with data.

Publishing then

One central place to receive your content

Publishing now



Heavy reliance on distribution networks

Business Problem

How can we understand our audience to...

Know what content to write

Optimize content in real time

Deliver content to the right user

Content engagement

How do people interact with content?

The Velocity Suite

Suite of products that empower the editorial team

Dash

Platform for seeing how content is performing across the Internet

Reports

Generate basic info about different topics, social media channels, or locations

CMS

In-house platform for writing and publishing articles

Kilogram

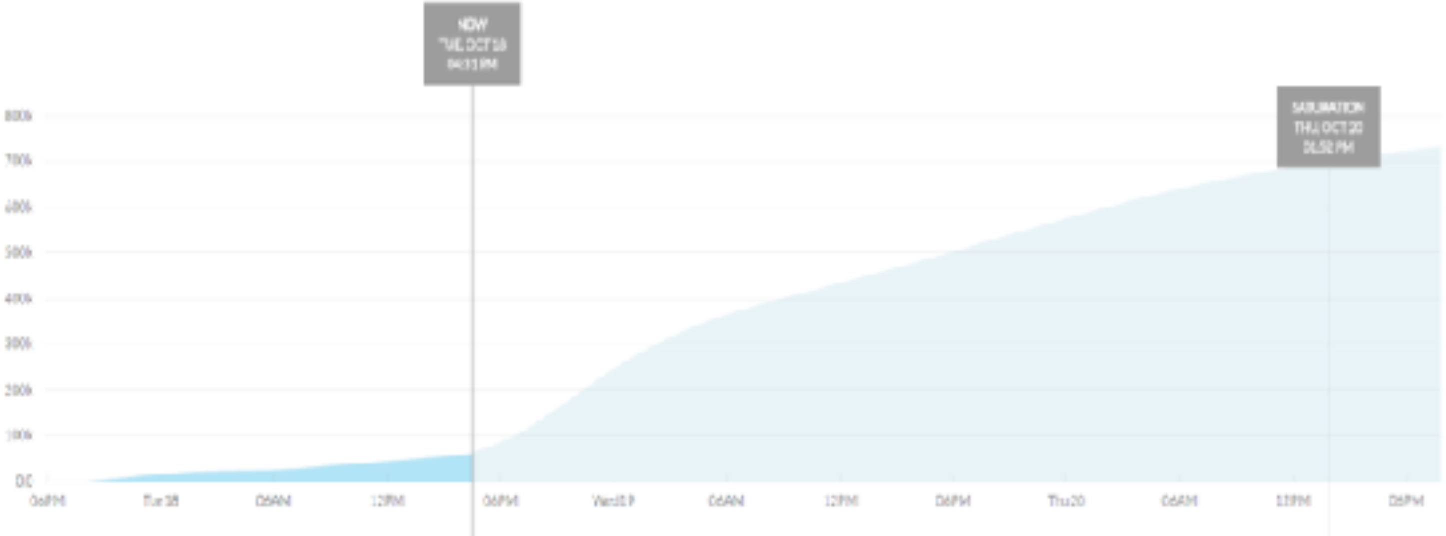
Understand how our content spreads across social media

Understanding Velocity

Crawl the web

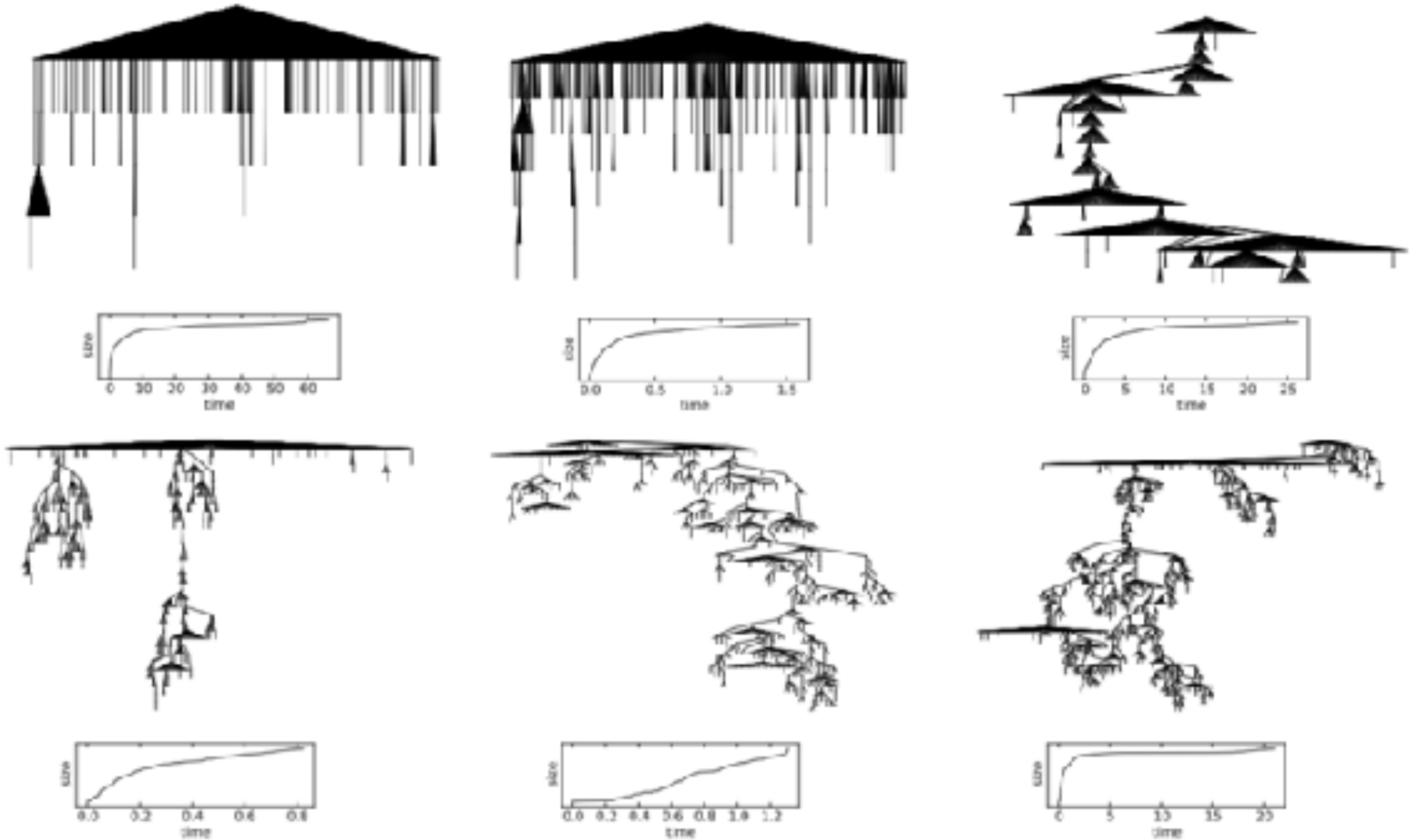
Track on social media

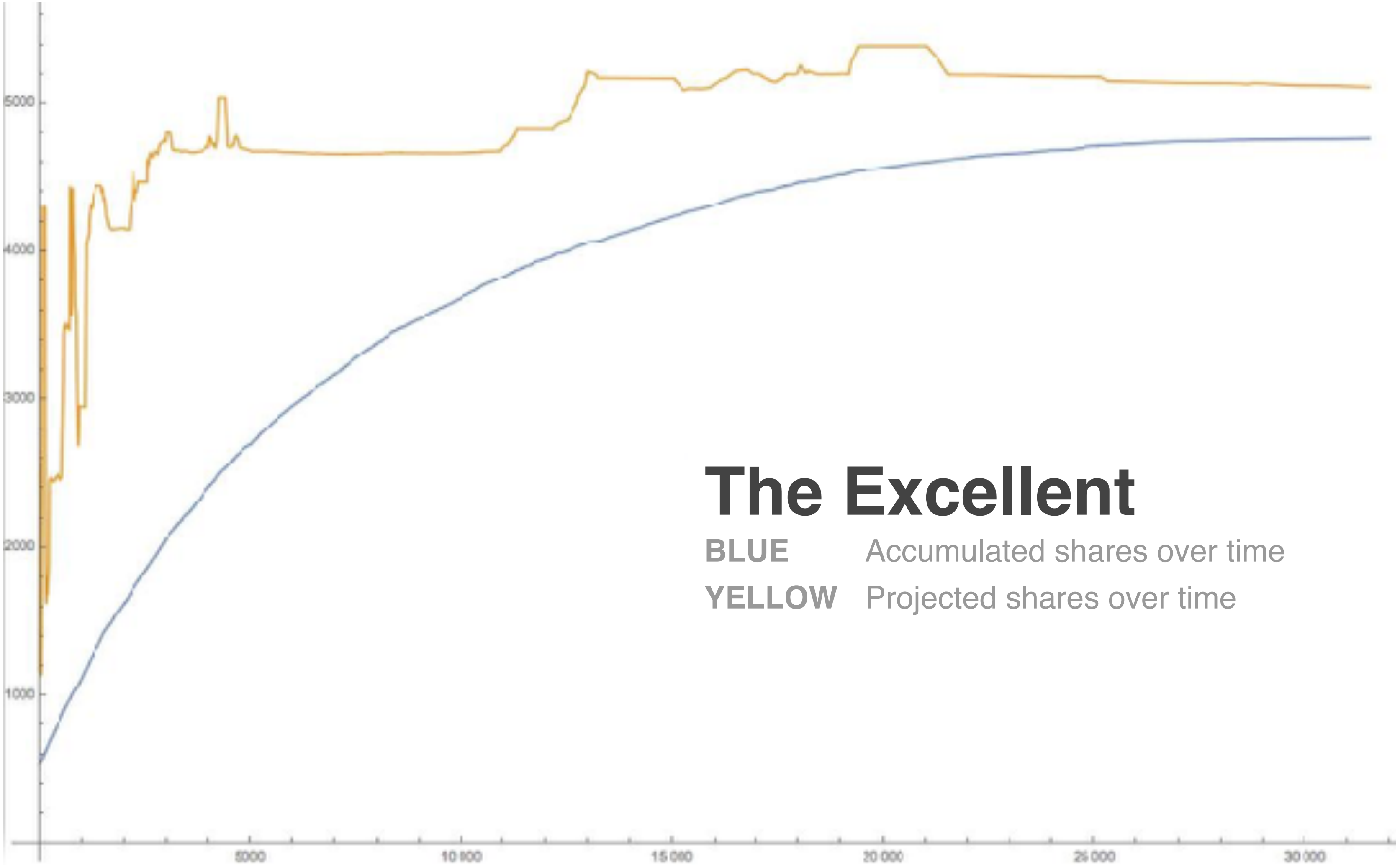
Predict future performance



The above is a cascade generate from a simulation with simple update rules. It bears a strong resemblance to what we actually see in our share button experiment. In fact, it turns out that a simple model of leaf growth/viewer rate yields a model of share behavior with predictive power!

$$S_t = \prod_{p=1}^{t} (1 + r_p \mathcal{N}_p) S_0$$

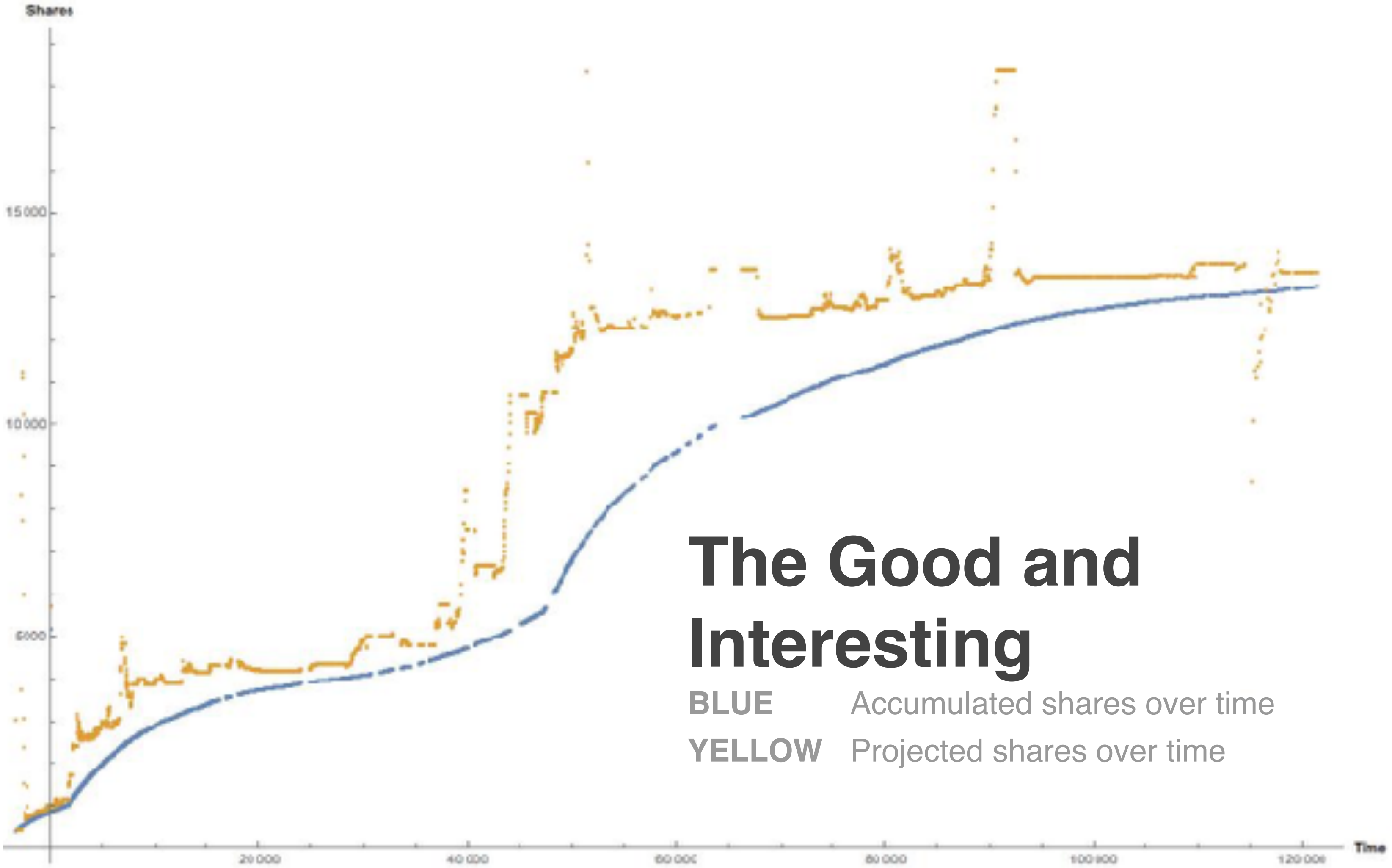

size
time
size
time
size
time
size
time
size
time
size
time



The Excellent

BLUE Accumulated shares over time

YELLOW Projected shares over time

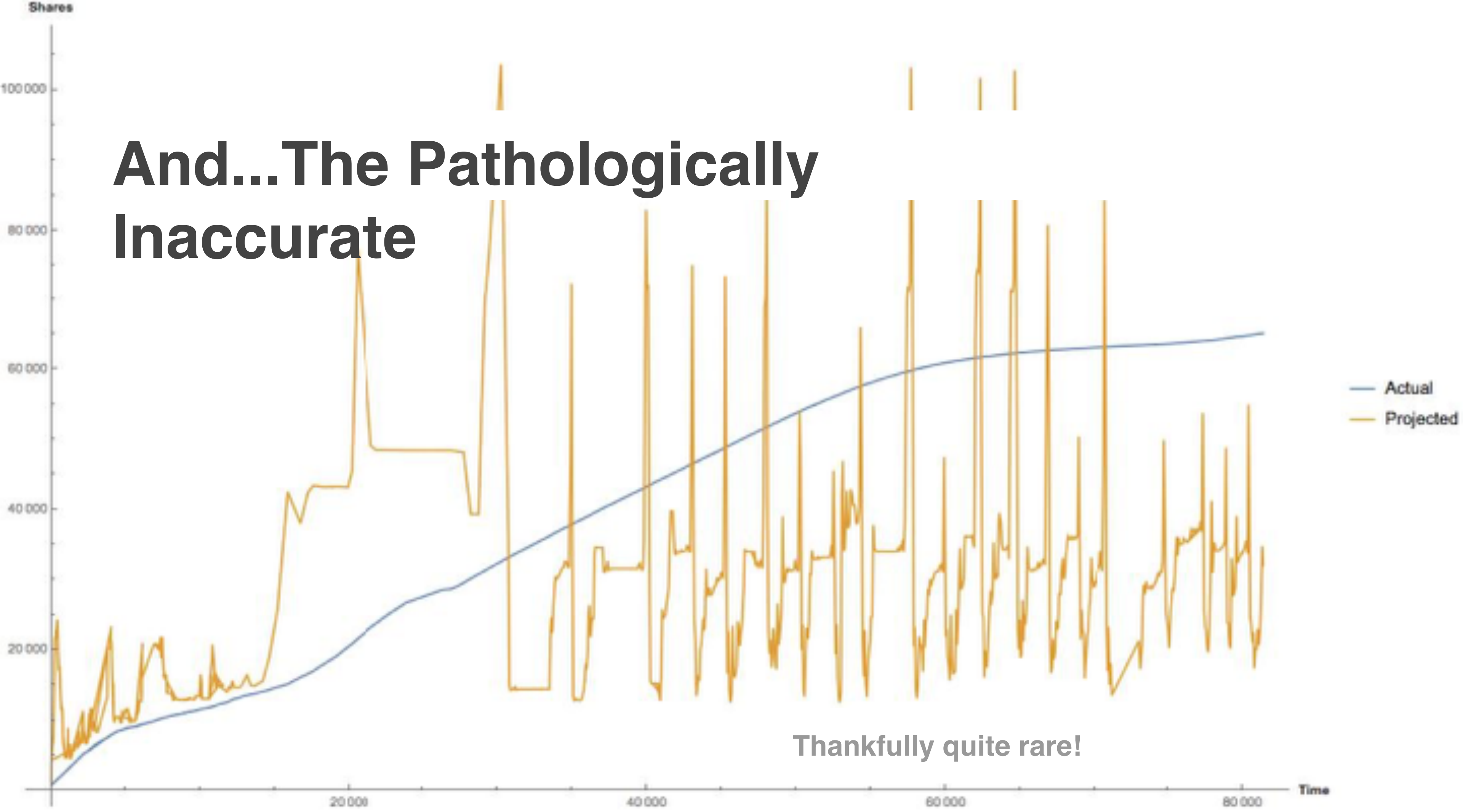


The Good and Interesting

BLUE Accumulated shares over time

YELLOW Projected shares over time

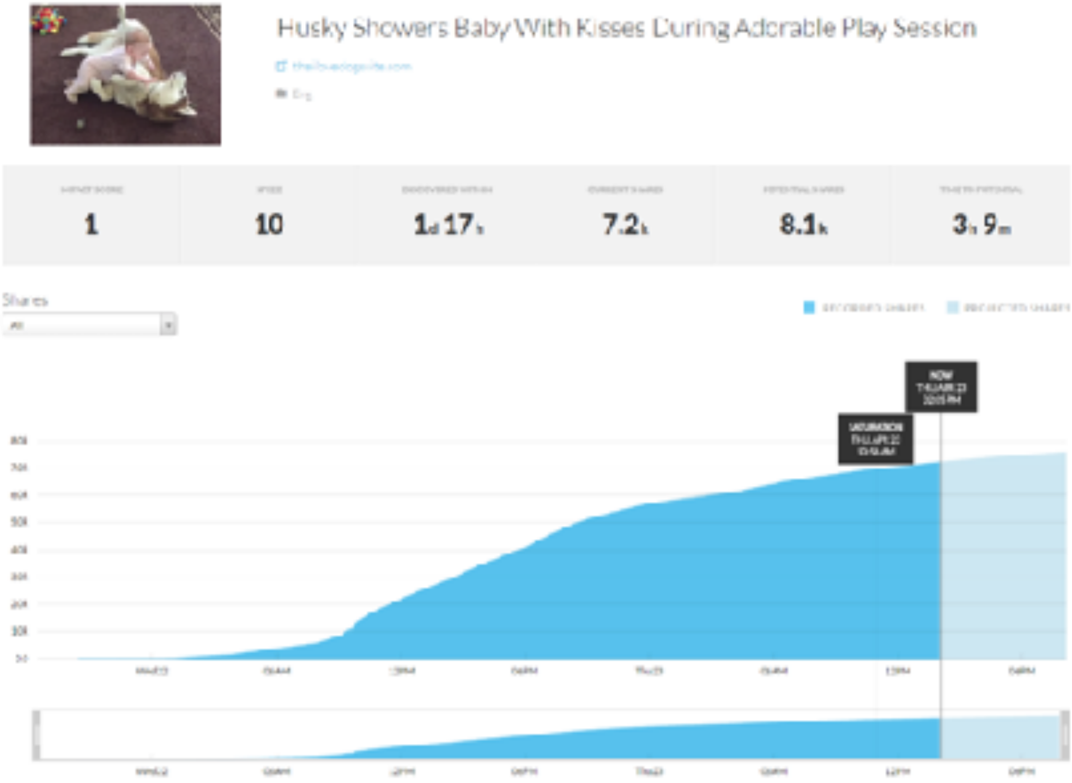
And...The Pathologically Inaccurate



Thankfully quite rare!

Velocity at work

- Discovered right as it was published
- Over 3000 data points collected
- Several points where trajectory changed & prediction found & adapted.
- Early projections very accurately modeled each subseries in the total dataset. Success!



Audience segmentation

What can we say about our audience?

Third-party segmentation?

Pros

- Out of the box
- Easy for non-technical stakeholders to interpret
- Can automatically import data from a variety of sources

Cons

- Not necessarily customized to our own audience, or what we're interested in
- Black box model – often unclear how segmentations are created

Demographic Data!

- Give editorial team profiles of users
- Identify similar users
- Analyze different sections of the site
- Use to drive content creation

	(Art & Theater Aficionados, 18-24, female)	(Art & Theater Aficionados, 18-24, male)	(Art & Theater Aficionados, 25-34, female)	(Art & Theater Aficionados, 25-34, male)	(Art & Theater Aficionados, 35-44, female)	(Art & Theater Aficionados, 35-44, male)
url						
/2016/07/01/game-of-thrones-season-7-predictions/	0.0	0.0	0.052364	0.0	0.0	0.0
/2016/07/01/game-of-thrones-theory-ned-stark-secret/	0.0	0.0	0.000000	0.0	0.0	0.0
/2016/07/01/kris-jenner-trolled-by-staples/	0.0	0.0	0.000000	0.0	0.0	0.0
/2016/07/01/queen-cersei-slams-boris-johnson/	0.0	0.0	0.000000	0.0	0.0	0.0
/2016/07/02/wrong-way-volte-face-photo-series/	0.0	0.0	0.000000	0.0	0.0	0.0

How should we model audience segmentation?

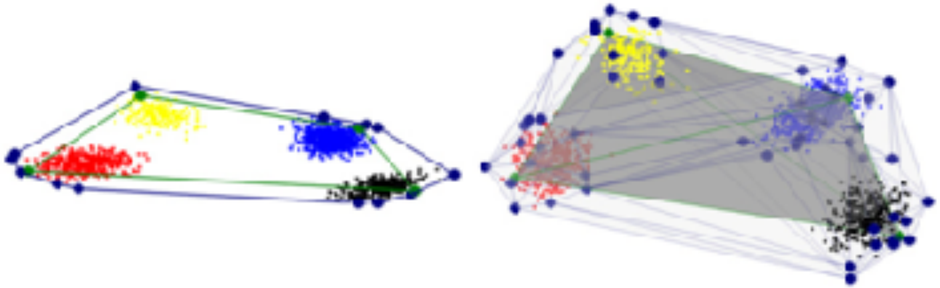
Option 1: Clustering

Option 2: Decompose the audience

Decomposition methods

Archetypal analysis

- Finds extremal points in multidimensional data as the basis for decomposition
- Archetypes are combinations of features

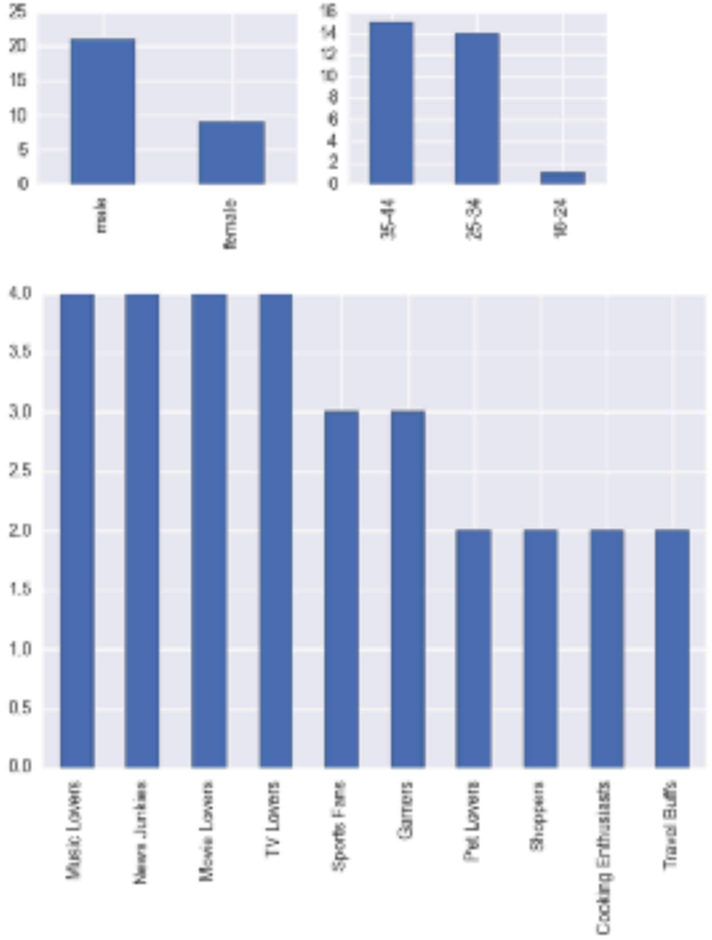


Non-negative matrix factorization

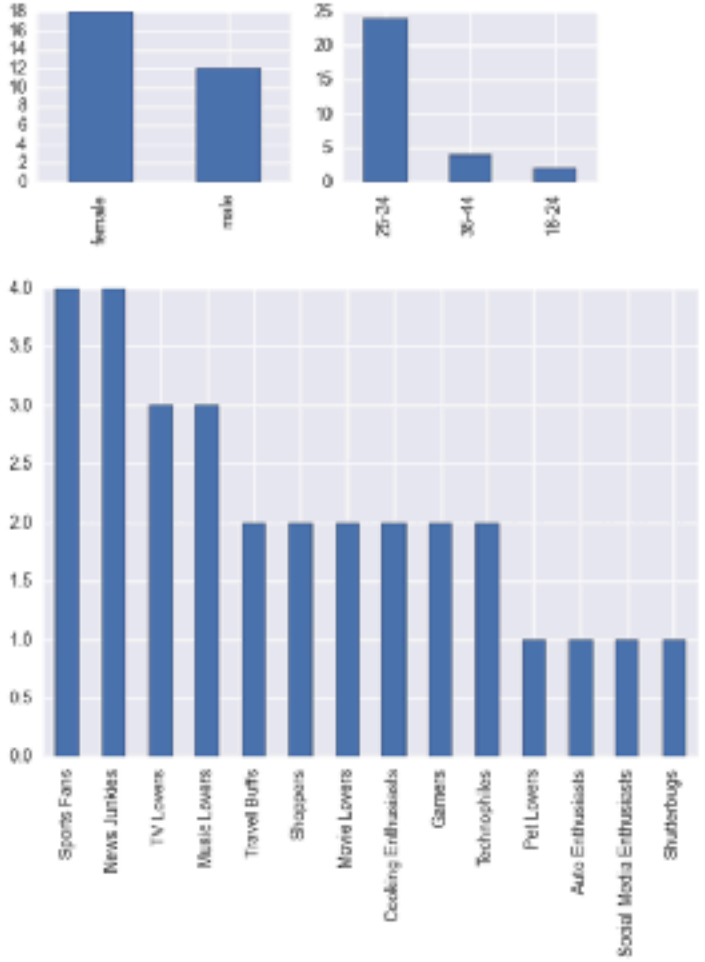
- Factors matrices in a way that allows for easier inspection
- Minimization of error function -> vector representation obtained in an additive fashion

Segment 1

NMF:

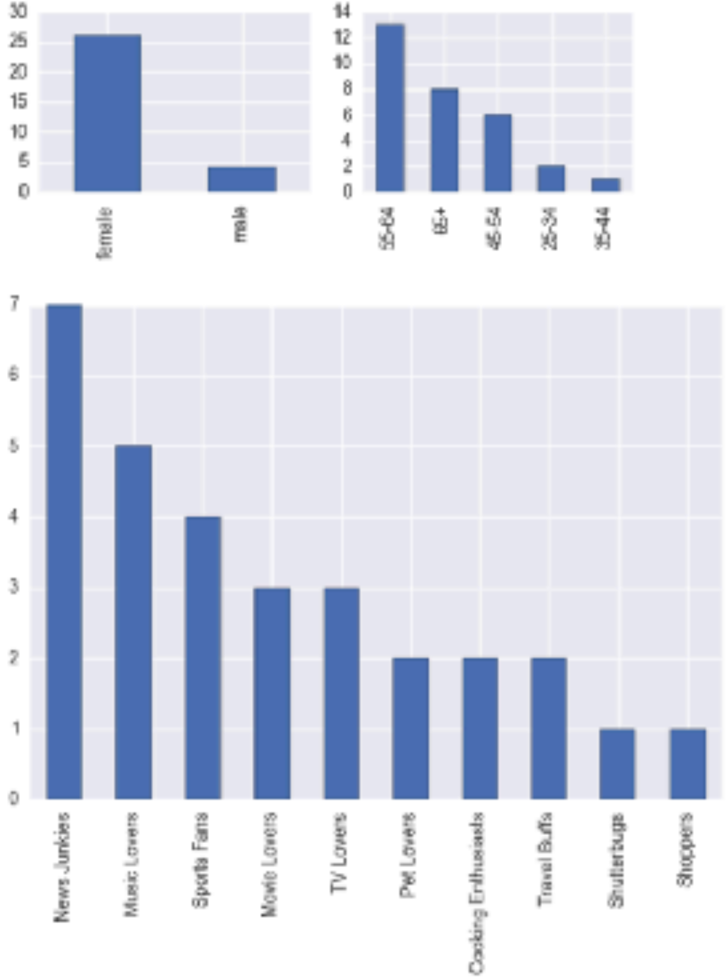


Archetypal:

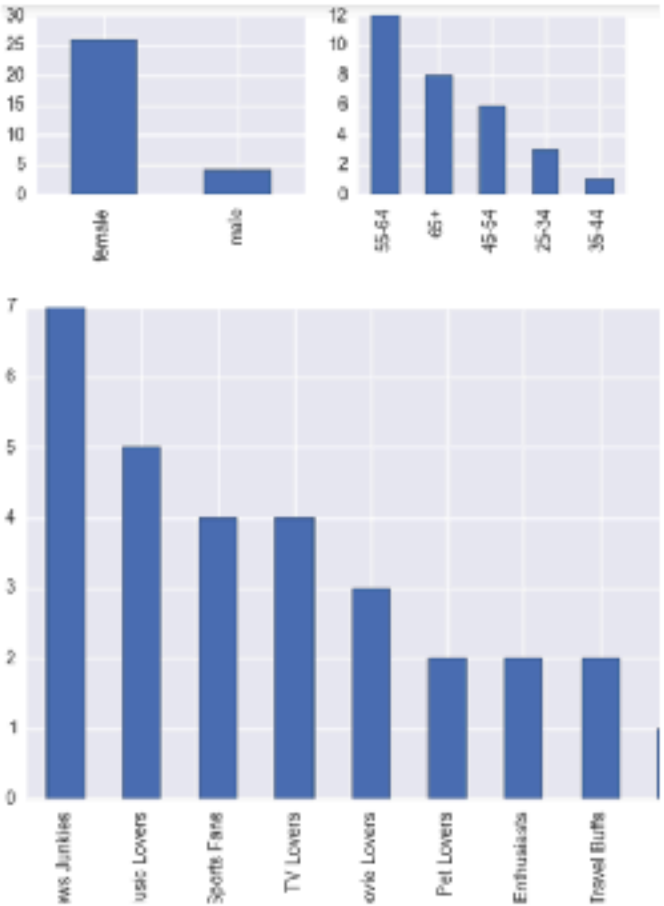


Segment 4

NMF:



Archetypal:



Caveats

- Google Analytics data – 1/3 of urls sent
- Finicky API
- Semi-useless interest data

Social media strategy

Where is our audience and how do we reach them?

Facebook Landscape

Multiple pages
Central “Main” page
Smaller secondary pages

Facebook Landscape

Multiple pages
Central “Main” page
Smaller secondary pages

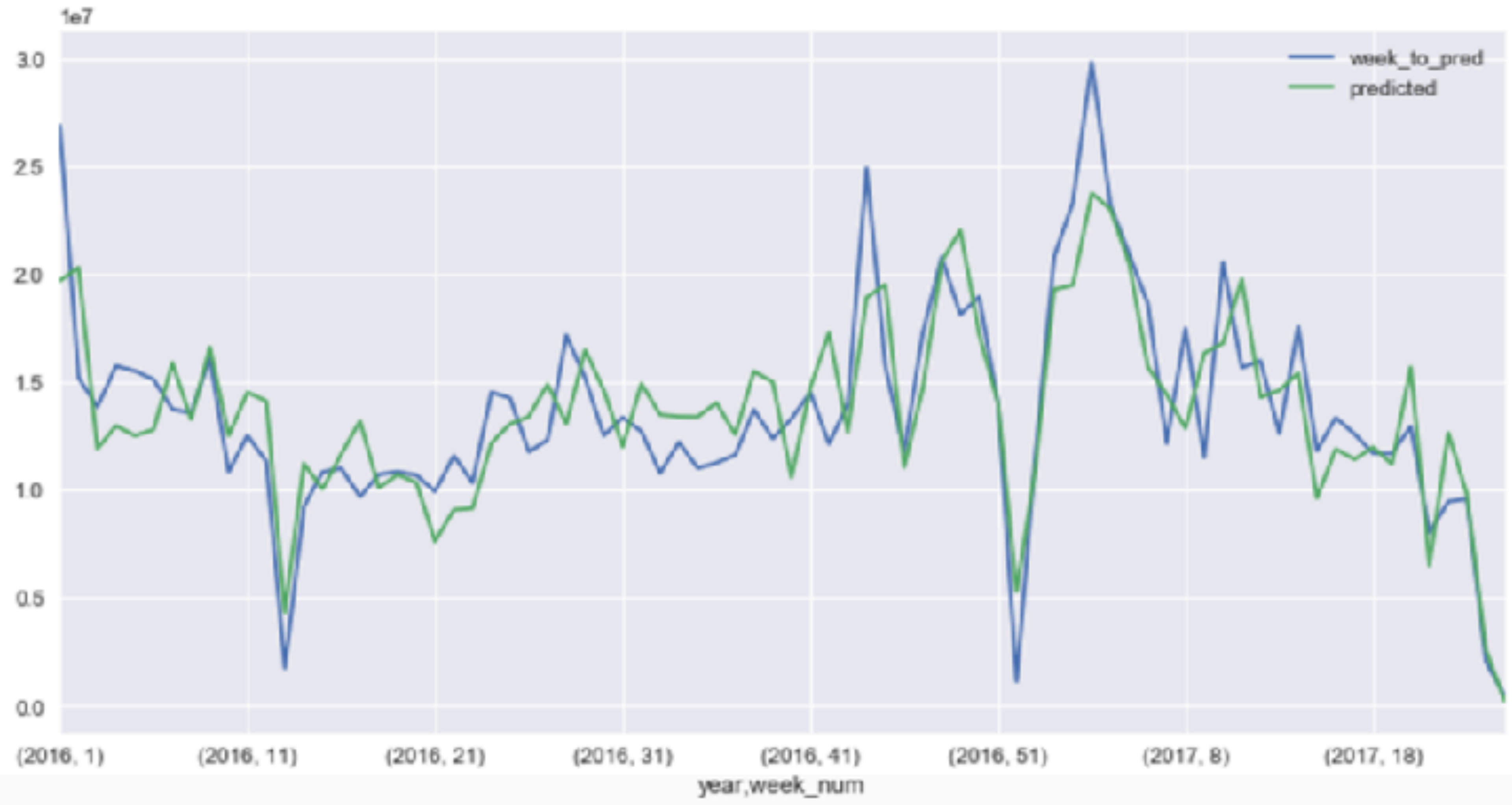
How do we develop an optimal Facebook strategy?

Is there a relationship between views and shares?

- Predict views from shares

	year	week_num	total_views	avg_engagements	url_count	total_engagements	week_to_pred
0	2016	1	20799986	1531.60206	872	1335557.00000	26938659.00000
1	2016	2	26938659	2118.85588	902	1911208.00000	15236620.00000
2	2016	3	15236620	1763.14508	772	1361148.00000	13872088.00000
3	2016	4	13872088	1410.99189	863	1217686.00000	15766583.00000
4	2016	5	15766583	1484.00118	851	1262885.00000	15538129.00000

- Linear regression
- Optimize for RMSE and MAE
  - RMSE: root mean square error
    - Standard deviation of residuals
    - Measure of spread in regression fit
  - MAE: mean absolute error
    - Average magnitude of errors
    - Less sensitive to larger errors


1e7
3.0
2.5
2.0
1.5
1.0
0.5
0.0
week_to_pred
predicted
(2016, 1)
(2016, 11)
(2016, 21)
(2016, 31)
(2016, 41)
(2016, 51)
(2017, 8)
(2017, 18)
year,week_num

How do we develop an optimal Facebook strategy?

Are there any decision points that are harming us?

Theory:
Once an article reaches 1k clicks on Twitter, we should post it to our main Facebook page.

Is this a good heuristic?

- Comparing populations – articles that achieved at least 1000 clicks on Twitter and were posted to main, and those that did not
- Are these populations different?
- What’s the performance of articles on our Facebook main page in each of these populations?
- Does using this heuristic perform better overall, in terms of views, than using no heuristic?

Population differences

- Population 1: articles that achieved 1000 clicks on Twitter and were posted to main page
- Population 2: articles that were not posted to the main page

- 2-sample Kolmogorov-Smirnov test
- Nonparametric test of equality of distributions
- Tells us that these two populations do NOT come from the same distribution

Gaussian process regression

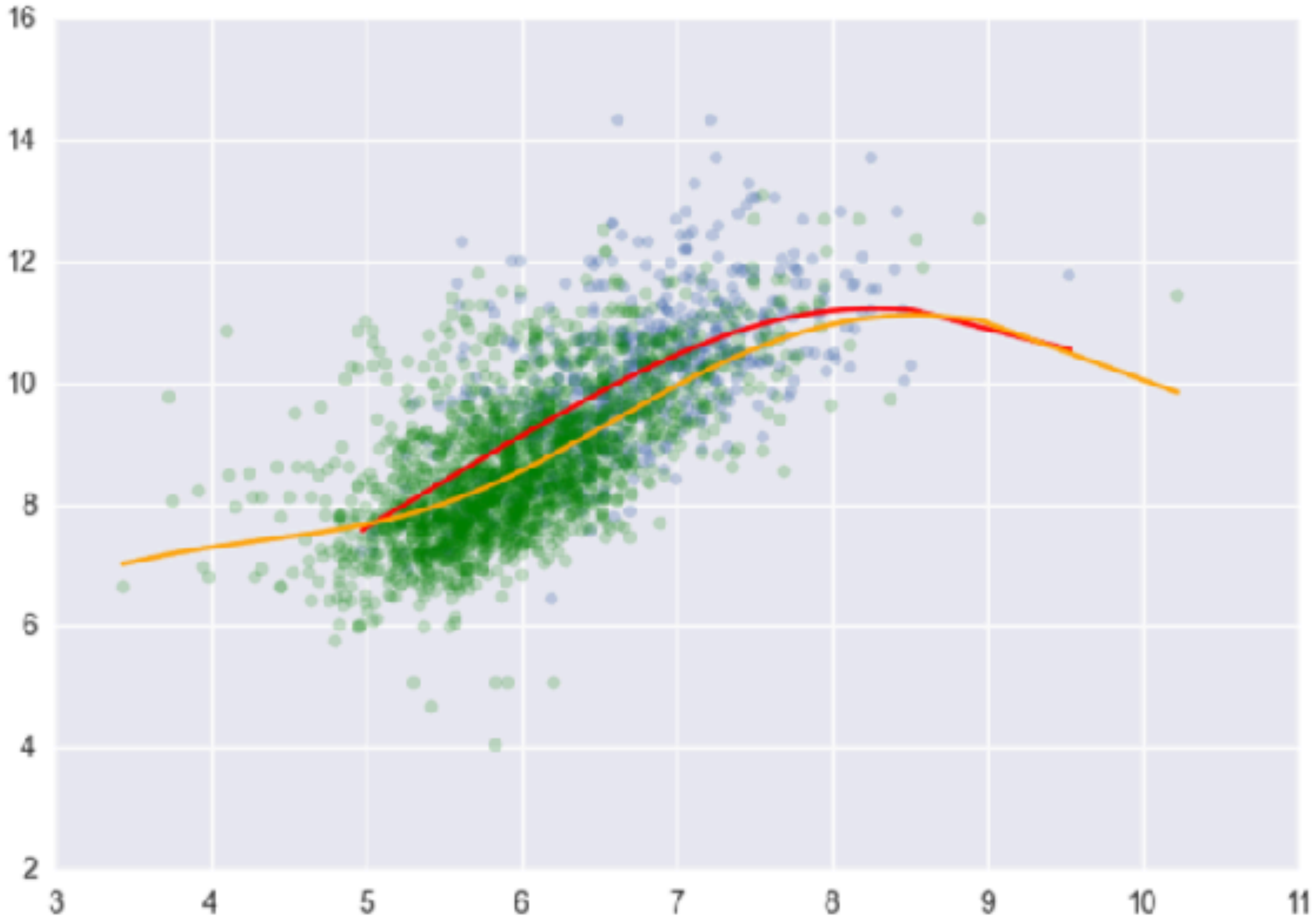
- Beyond linear regression
- Nonparametric approach to finding the distribution over all possible functions f(x) that are consistent with observed data

Why?

- Gives us a full conditional distribution
- Probability that an article will achieve *n* page views both using the heuristic and not

How?

- Build two Gaussian process regression models: one on the views in the “main” set, one on the views in the “not main” set and use to simulate what could happen


16
14
12
10
8
6
4
2
3
4
5
6
7
8
9
10
11

Method

1. Build the models for each population
2. Sample a large number of times at Twitter clicks = 1000 for each Gaussian process regression and make a views prediction
3. Determine how frequently a “main” sample has higher page views than a “not main” sample by drawing from each sample a large number of times
4. Determine how frequently a randomly chosen “main” sample beats a “not main” sample, without using any heuristic.

Result: Using the 1000 clicks on Twitter heuristic, posting to the main page achieves higher page views than not posting to the main page **65%** of the time.

Result: Using the 1000 clicks on Twitter heuristic, posting to the main page achieves higher page views than not posting to the main page **65%** of the time.

Result: By selecting **randomly** and using no heuristic, posting to the main page achieves higher page views than not posting to the main page **78%** of the time.

Final thoughts and wrap-up

Takeaways

- Audience data can be messy and complex
- Make data usable for nontechnical stakeholders
- Have an understanding of both the audience reading your content and the audience you're developing for
- Know what metrics you want to optimize for
- Know what your end goal is
- Optimize for interpretability

Empower our editorial team through data, not with data

Current and future work

- Facebook Index
- Velocity 2.0
- Behavioral analysis of session data
- Headline optimization
- All things video

Papers and blog posts

- *The structural virality of online diffusion*. Goel, Anderson, Hofman, Watts. 2015.
- *Archetypal analysis*. Cutler, Breiman. 1994.
- https://github.com/ulfaslak/py_pcha, Python archetypes package.
- http://katbailey.github.io/post/gaussian-processes-for-dummies/, Katherine Bailey
- https://blog.dominodatalab.com/fitting-gaussian-process-models-python/, Chris Fonnesbeck
- Scikit-learn documentation!

Thank you!

Sarah Guido

Senior Data Scientist

Mashable

@sarah_guido

Please

Remember to rate this session

Thank you!

@GOTOber

gotober.com