The Next Big Sound? Modeling the Path from Relative Obscurity to the Top of the Charts Using Social Metrics

Rebecca Nugent
Department of Statistics, Carnegie Mellon University
joint with Emily Wright

http://www.stat.cmu.edu/∼rnugent/PCMI2016

PCMI Undergraduate Summer School 2016

July 19, 2016

Music Analytics: Spreading the Word

Music Genome Project:

- tracks up to 400 attributes per song (genre, vocalist gender, distortion on guitar, key tonality, instrument proficiency, etc)
- complete list of attributes is proprietary/trade secret
- organizes/archives songs with mathematical algorithms
- uses the "distance" between songs to make recommendations
- Pandora Internet Radio: music streaming and automatic recommendation; can give positive/negative feedback
- licensing only for certain countries;
  can buy subscription to skip ads; can't rewind or repeat

Identifies songs from a large database based on user-defined characteristics; songs are pre-analyzed; algorithm updates/adapts

Music Analytics: Spreading the Word

Spotify/Grooveshark:

- commercial music-streaming; about 15-20 million songs
- can browse by artist, album, genre, etc; can purchase
- can create/share playlists
- allows creation of random playlist with specific characteristics
- can rate songs to help improve playlist creation
- several active lawsuits (Grooveshark)

Music library that user can browse and "borrow"; recommendations based on song "categories", not mathematical features

Criticism includes inadequate royalties, copyright infringement; attention drawn to how little new artists get paid

How does it help the artists?

- potentially introduces them to users
- apparently doesn't pay them that much

Artists only "discovered" if someone stumbles upon them

Music Analytics: Spreading the Word

New artists turned to social media platforms (before record labels did) in an effort to build a fan base

- Myspace: social networking service; now primarily focused on music (co-owned by Justin Timberlake)
- Youtube: video-sharing website; users can start their own "channel', have subscribers
  - Justin Bieber: discovered by Scooter Braun
  - Avery, Alyssa Bernal, Greyson Chance, Dondria
- Soundcloud: artists can create and upload to a unique url; can be combined with Twitter/Facebook
- The Hype Machine: compilation of blogs about music; can browse for new artists

Music Analytics: Spreading the Word

Next Big Sound: 1st to delve into analytics for prediction

- started as an assignment in an entrepreneur class at Northwestern University; $25,000 in venture capital
- named one of the 10 best music startups in 2010 (Billboard magazine); CEO named "one to watch"
- tracks artists by their social metrics and fan base demographics: NBS Profile Features

"Predicting the Next Big Sound": weekly list of 15 artists poised to breakthrough based on their social trends

A large, searchable database on 100,000s of artists;
can browse to find information about artist of interest;
can subscribe to reports; music industry can track artists;
need to actively use/belong to discover new artists

Music Analytics: Spreading the Word

Radio use (including online/satellite) still constant

% of Americans (12 or older) who Use/Own Platform or Device

	2001	2009	2010	2011
Television	98	NA	98	98
Local AM/FM Radio	96	92	92	93
Cellphone	54	81	84	84
Broadband Internet	20	NA	64	70
Online Radio	28	49	52	56
Online Video	23	NA	49	54
Facebook	NA	NA	48	51
YouDVR	NA	NA	41	36
Tube	NA	NA	46	49
VoD	20	NA	NA	35
iPod	NA	28	28	31
Smartphone	NA	NA	NA	31
Audio Podcasts	NA	22	23	25

stateofthemedia.org; Arbitron

Getting on radio is still the goal

(*That Thing You Do*, 20th Century Fox, 1996)

How Do New Artists Get on Radio?

Goal: Determine Associations over Time between Social Media Usage and Heavy Radio Airplay for Newer Artists

Steps:

- New artists predicted to be big: Next Big Sound daily metrics
- Radio airplay data - which artists have made it to the top?
- Link these two databases together: **statistical record linkage**
- Generate variables that describe/summarize our time series; **classification model** for whether artists end up on radio (treats observations like they're independent)
- **Impute** missing values for social media metrics
- **Longitudinal models** to predict artist "survival" over time

Radio Airplay Data

- Weekly airplay data from over 5000 terrestrial radio and online channels; Digital Radio Tracker; Webscraped using Excel
- Scrape the "Top XX" Charts for 2013
  (Top 200, Top 50 Pop, Top 50 Country, etc)
- Seven fields for each song: artist's rank, artist name, song, weekly airplay, month, day, year; 19,550 observations

Rank	Artist	Title	Airplay	Month	Day	Year
1	Rihanna	Diamonds	5768	1	5	2013
2	Bruno Mars	Locked Out of Heaven	4500	1	5	2013
3	Flo Rida	I Cry	3765	1	5	2013
4	Maroon 5	One More Night	3339	1	5	2013
5	Ke$ha	Die Young	2875	1	5	2013
6	fun.	Some Nights	2629	1	5	2013
8	P!nk	Try	2124	1	5	2013
35	Smoke	I Ain't Hiding (w/ T-Pain)	1167	1	5	2013
123	Meka Arpege & Barshaun	Keys of Hope	707	1	19	2013

Radio Airplay Data

Eventually need to link this data with the Next Big Sound metrics

Issues include:

- no artist IDs
- text strings vary (accents, typos, punctuation, etc)
- no uniform listing for featured artists
  sometimes in title, sometimes in artist; "featuring", "featured", "ft", "with", "w/", "and"
- Required text string parsing: automatic, human spot-checked
  "Beyonce and Lady Gaga" - two artists
  "Florence and The Machine" - one artist
- each artist now gets credit for the song

For this project, "on radio" means listed on a Top XX chart in 2013

Next Big Sound Predicted Artists

Generous collaboration with Next Big Sound
Access to data; not access to their prediction algorithms

- Weekly lists of 15 artists predicted to be "big";
  have the fastest accelerating online activity
- August 2010 to December 2013
- Scraped artist lists first using R/XML/Curl (D. Lang)
- List includes Next Big Sound ID, artist's rank, artist's name
- 3,213 artists total
- post-processing included:
  - de-duplicating artists (appeared on more than one list);
    kept the most recent
  - resolving artists with same name but two different NBS IDs
    (DJ Drama, Mavado)

Next Big Sound Predicted Artists

For each artist, we collect daily social metric data from NBS

- Java program (Kelvin Rojas, CMU)
- Uses NBS ID to extract artist profile to an XML file
- First pass: key social media metrics
  - Facebook page likes
  - Wikipedia page views
  - Twitter followers
  - YouTube views
  - Vevo plays
  - Soundcloud plays
- Metrics are total values to date;
  don't always have all metrics for all four years

Of these over 3000 artists, how do we know which ones were successful on radio? (top of the charts?)

Need to **link** list of NBS artist records to list of radio records

Record Linkage for NBS and Radio

Match records of unique individuals across two data sources

Primarily a comparison of text strings and numeric values

Can be algorithm or statistical model, supervised or unsupervised; used to disambiguate/de-duplicate

Examples:

- Linking Census Records to other surveys
- Determining author/inventor ownership (bibliometric records/patents)
- Linking information on homicides in Colombia
- Determining the civilian casualty count in the Syrian civil war

Usually results in a probability of matching for each pair of records

Record Linkage for NBS and Radio

Simple, One Field Comparison: compare NSB name to radio name
Why can't we just do exact string matching?
Typos, punctuation, capital letters, etc;
instead we use a *similarity metric*

- Jaro-Winkler: looks at transpositions needed to correct typos

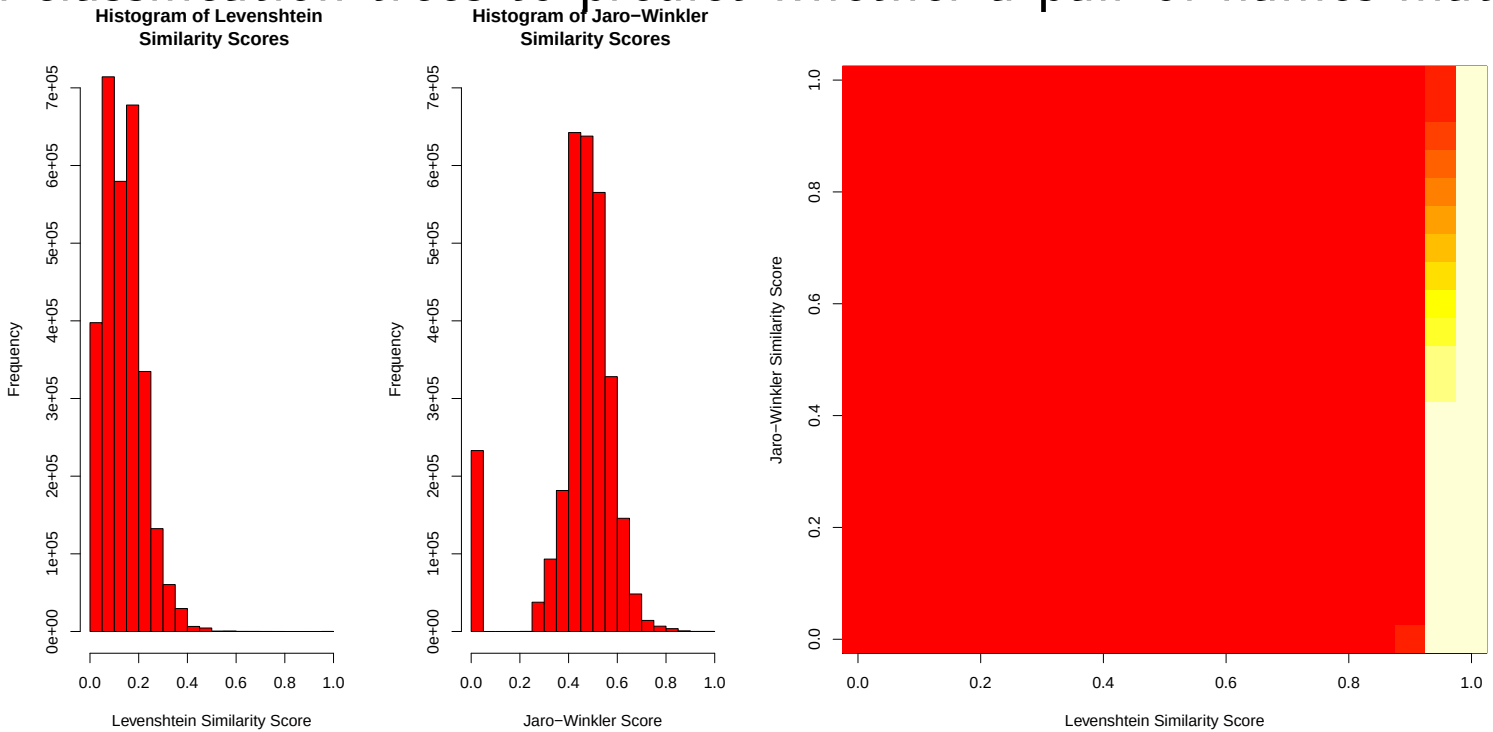
$$\phi_J(s_1, s_2) = W_1 \cdot \frac{c}{L_1} + W_2 \cdot \frac{c}{L_2} + W_t \cdot \frac{c - \tau}{c}$$

- Levenshteim: transformed edit distance; edits are insertions, missions, substitutions

NBS Artist	Radio Artist	Lev.	JW
doble man	hrc	0	0
dayan	keith urban	0.182	0
blessed by a burden	joni mitchell	0.053	0.253
the popopopops	3 doors down	0.214	0.540
lance herbstrong	lana del rey	0.375	0.705
deuce	deuce.d	0.714	0.943
atlas genius	atlas genius	1	1

Predicting Matches

First build small hand-matched data set with pairs of names that match and don't match (n = 640); use *logistic regression models* and *classification trees* to predict whether a pair of names match.



In both cases, JW score was the important once (p = 0.09; split twice in tree); classified only a few pairs incorrectly

ty dolla $ign vs ty dolla fun. vs fun

Using Social Media Metrics to Predict Success on Radio

Now we know which NBS artists have made it to a "top chart"; we then turn to their social media metrics:

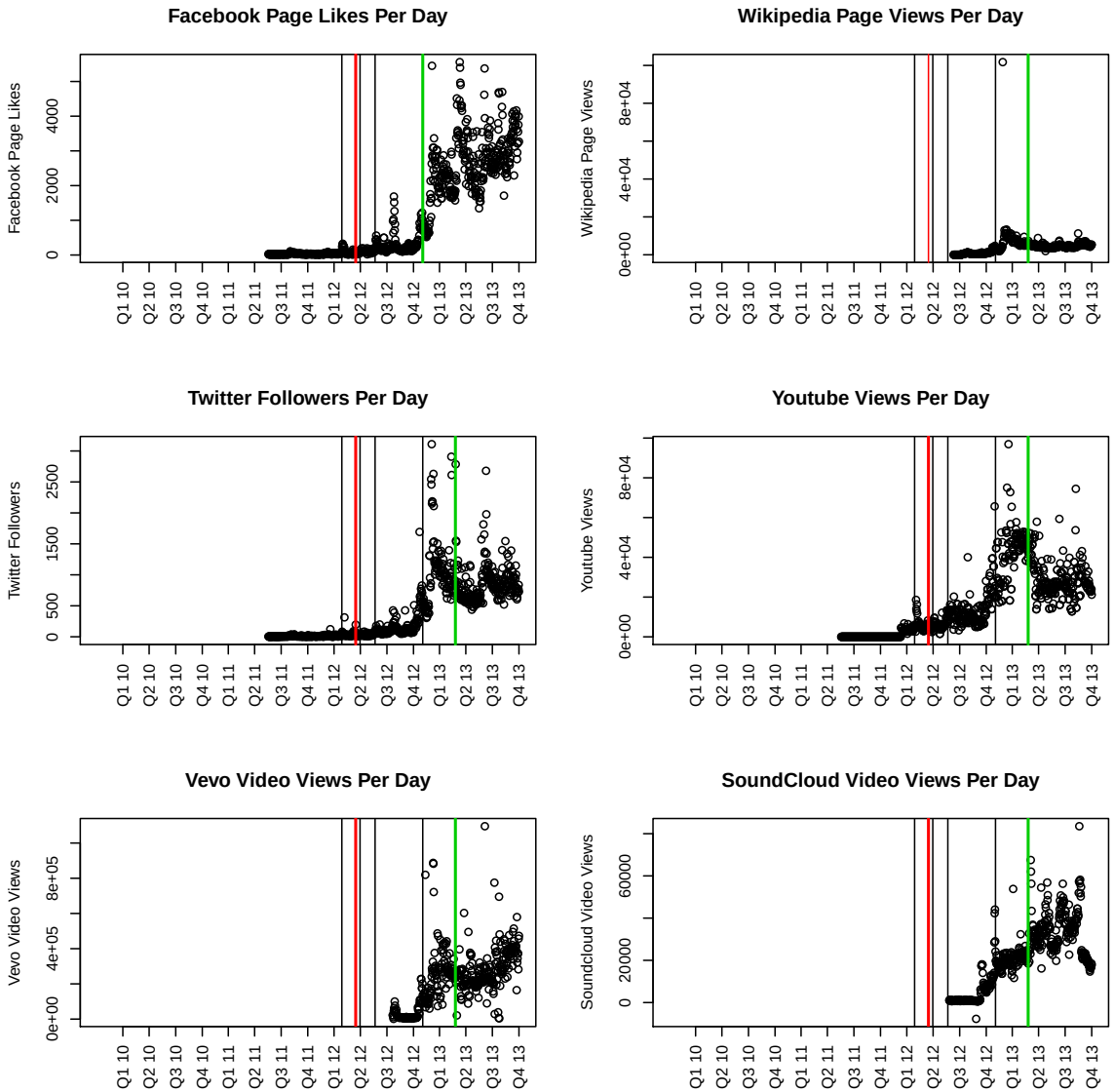
Bastille:

Table 13: Summary Statistics of Bastille Net Daily Online Metrics				
	Min	Max	Mean	Std. Dev
Facebook Page Likes	7	5,558	1,089	45.9
Wikipedia Page Views	0	101,770	4,156	238.5
Twitter Followers	-1	3,110	395.2	17.0
Youtube Video Views	0	96,920	16,110	562.4
Vevo Video Views	1,277	1,096,000	217,200	7,539.0
SoundCloud Plays	7,727	83,530	21,130	684.8

April 27, 2012: First song release "Overjoyed"; June 14, 2012: Predicted by Next Big Sound; June 29, 2012: Vevo Video of "Bad Blood" released; Aug 20, 2012: "Bad Blood" digitally releases; Feb 2013: "Pompeii" song released; May 25, 2013: First appearance on Top 200 radio song chart

Social Media Metrics

Bastille:



Using Social Media Metrics to Predict Success on Radio

Big Boi:

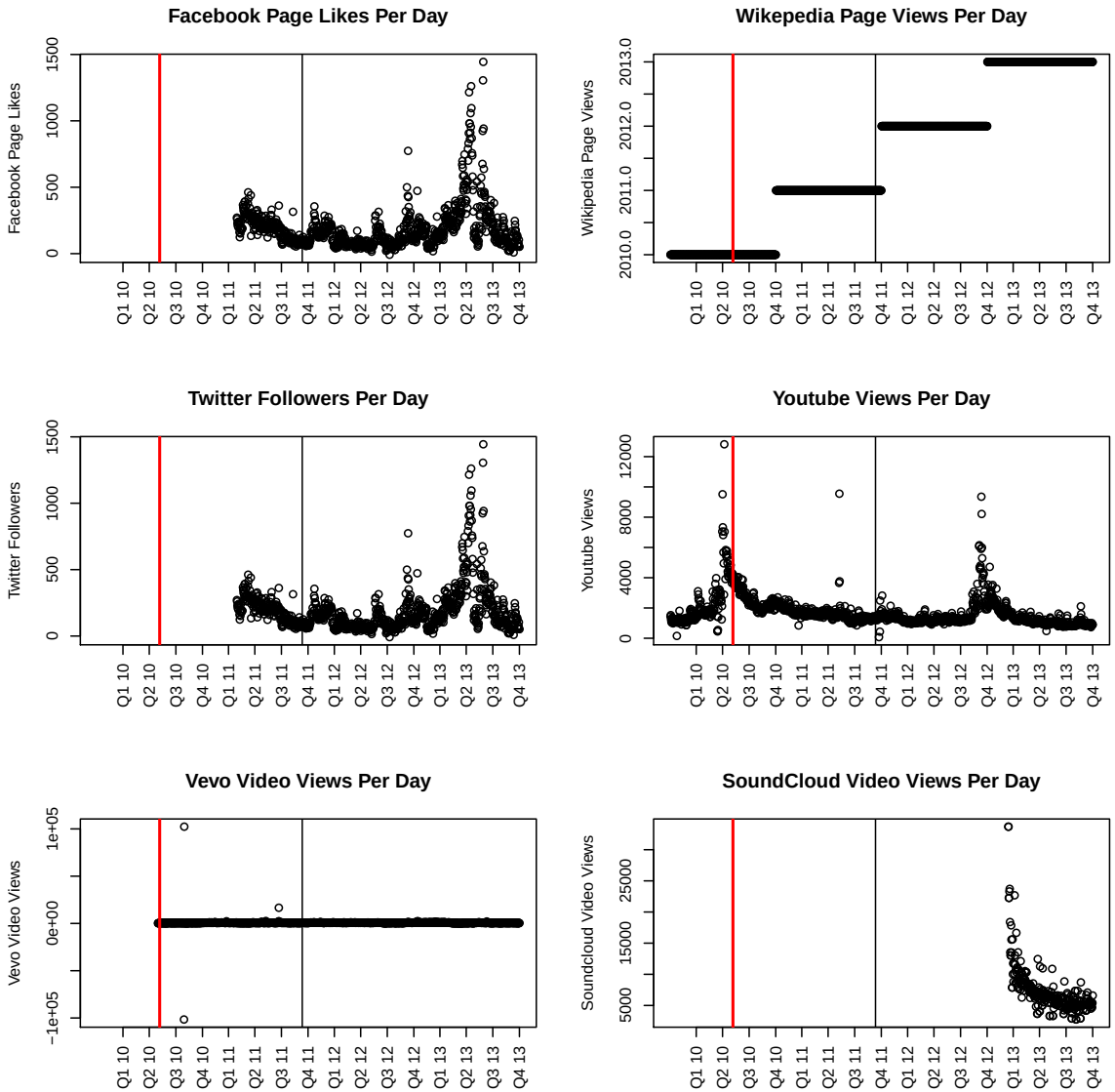
Table 14: Summary Statistics of Big Boi Net Daily Online Metrics				
	Min	Max	Mean	Std. Dev
Facebook Page Likes	-8	1,444	187.4	5.2
Wikipedia Page Views	75	12,820	1,710	26.2
Twitter Followers	-101,600	102,400	670.7	116.3
Youtube Video Views	2,746	33,700	7,335	234.8
Vevo Video Views	125	1,537,000	17,630	1,596.3
SoundCloud Plays	20	236,900	5,907	322.4

Aug 5, 2010: Predicted by Next Big Sound;

Dec 11, 2011: Release of second album "Vicious and Dangerous Rumors"

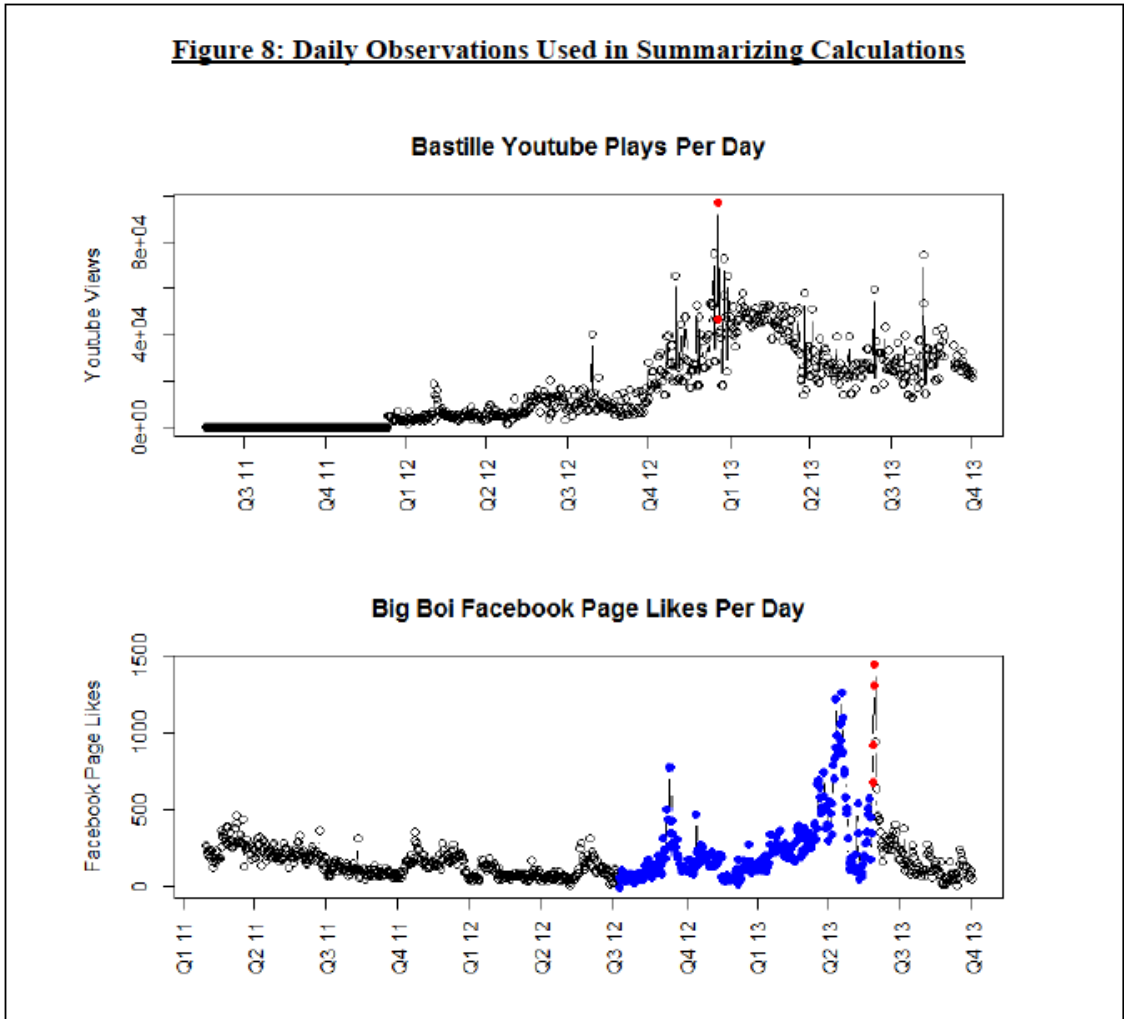
Using Social Media Metrics to Predict Success on Radio

Big Boi:



Using Social Media Metrics to Predict Success on Radio

Also used: average value over days, weeks, months; maximum net daily value; aggregate before peak value; sum of daily value from beginning to maximum; peak slope; percentage increase over time



Using Social Media Metrics to Predict Success on Radio

Starting with just a logistic regression predicting whether or not the NBS predicted artist made it to a "top 'chart"

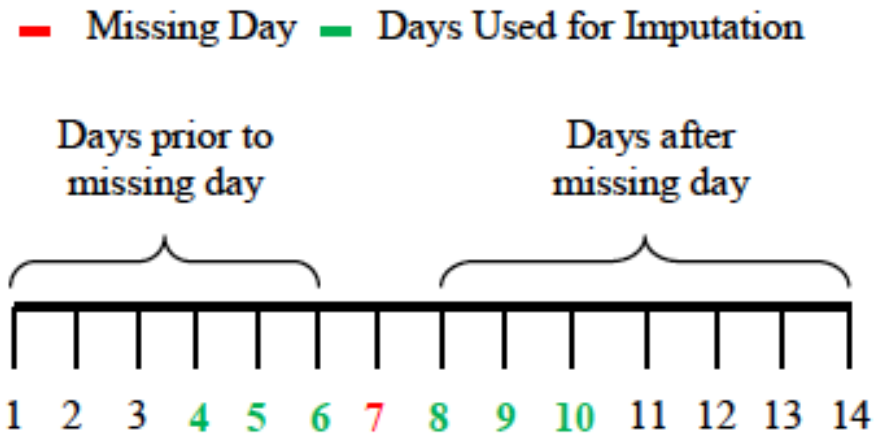
Significant Predictors:

- Average daily Twitter followers (negative) and Soundcloud use (positive)
- Facebook maximum increase (positive)
- Facebook Aggregate before Peak A (negative)
- Soundcloud Aggregate before Peak C (negative)
- Facebook Peak Slope (negative)
- Facebook Percentage Change Over Time (positive)

Current Imputation Attempts

Missing values don't allow us to put in all metrics for all days into the model (Vevo, Soundcloud)

Can use imputation techniques to try to make "good guesses"
Exploring averaging over ranges of 1-3 days before, 1-3 days after:



- Does a great job for metrics like Facebook (within 2%)

  using more previous info tends to underestimate;
  using more later info tends to overestimate
- Does a bad job for metrics like Wikipedia views (within 40%)

We impute missing metrics only if there is at least one value within +/- 3 days; otherwise, stays missing

Predicting "Survival" over time

Metrics are time series; treating as independent not a good idea. Instead we take a look at Cox Proportional Hazard Models:

$h(t)$ is the estimated risk at time t

$h_0(t)$ is the baseline risk at time t which is solely dependent on time

p is the number of covariates

β_j is the estimated coefficient and hence size of effect for the j^{th} covariate

$$h(t) = h_0(t) * e^{\sum_{j=1}^{p} \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots \beta_p x_p} \quad (1)$$

$$\frac{h(t)}{h_0(t)} = e^{\sum_{j=1}^{p} \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots \beta_p x_p} \quad (2)$$

$$\log\left(\frac{h(t)}{h_0(t)}\right) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p \quad (3)$$

Survival Model Results

Modeling "Survival" over time; observations followed until "death"
Here, Death is making it on a "top chart"

Table 26: Multivariate Cox Proportional Hazards Model - All Weeks

22 Radio Artist, 60 Total Artists

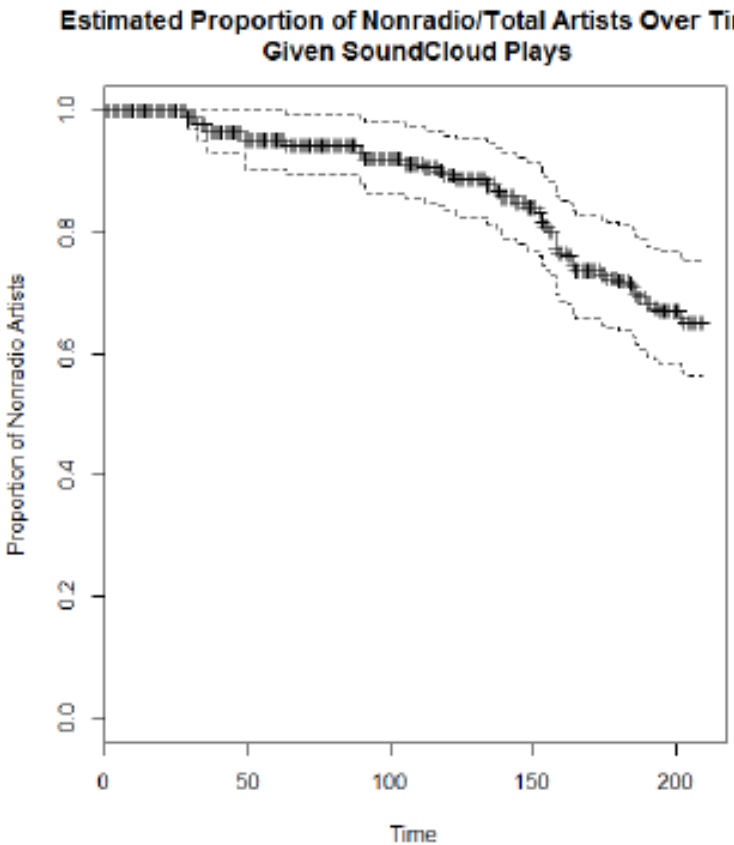
Global Proportion Hazard Assumption = 0.85

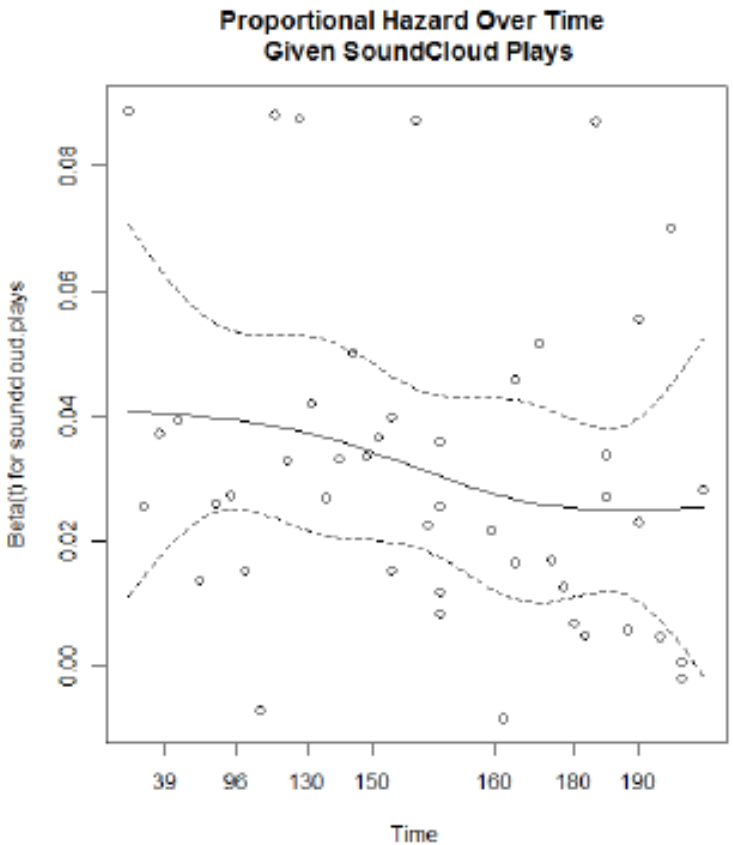
Bolded models are statistically significant at the 1% level

Variable	Coefficient	Exponential of Coefficient	P-Value	Proportional Hazard Violation P-Value
Facebook Page Likes	-0.027	0.973	0.33	0.64
Wikipedia Page Views	-0.0053	0.995	0.40	0.35
Twitter Followers	0.031	1.031	0.13	0.53
Youtube Plays	0.00045	1.00045	0.13	0.93
Vevo Plays	0.00073	1.00073	0.12	0.49
SoundCloud Plays	**0.0205**	**1.0207**	**0.0047**	**0.73**

Survival Model Results

Closer look at SoundCloud





Future Work/Conclusions:

- Incorporate more metrics (sub-metrics)
- Work with NBS to develop prediction algorithm
- Include information akin to Music Genome Project (One Million Song Project??)

http://www.stat.cmu.edu/∼rnugent
rnugent@stat.cmu.edu