A Day in the Life of a Data Scientist

Brian Eoff (@bde)
Lead Scientist at Spring

SPRING

What is data science?

Jake Vanderplas
@jakevdp

Follow

Best thing about being a data scientist: nobody actually knows what data science *is*, so you can pretty much do whatever you want.

RETWEETS 65 FAVORITES 72

9:28 AM - 17 Oct 2014

Data Science Conference Bingo Card

In-Memory	Unstructured Data	"We're Hiring"	Predictive Analytics	Streaming
Iris Data Set	Machine Learning	Real Time	Datafication	Facebook and Twitter
NoSQL	Mobile	Free Space!!	Internet of Things	Reuters-21578
Visualization	Hadoop	Social Graph	@BigDataBorat quote	Wordcount Demo
Sentiment Analysis	NCDC GSOD	Business Intelligence	Someone who thinks R doesn't suck	"Data Is The New Oil"

Vinod Khosla: In The Next 10 Years, Data Science Will Do More For Medicine Than All Biological Sciences Combined

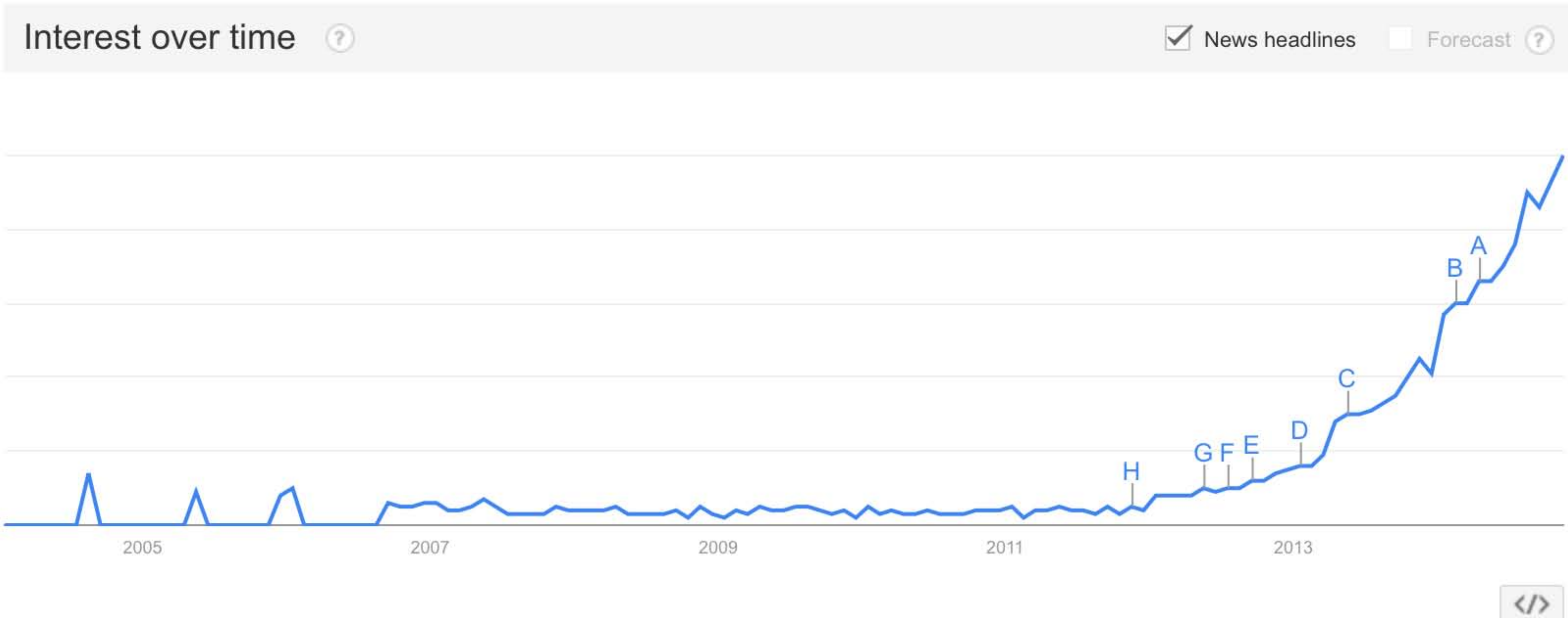
FREDERIC LARDINOIS

Wednesday, September 11th, 2013

Comments

Data Scientist: The Sexiest Job of the 21st Century

by Thomas H. Davenport and D.J. Patil



Interest over time
News headlines
Forecast
A
B
C
D
E
F
G
H
2005
2007
2009
2011
2013

2011 May - Join Bitly as Data Scientist

2013 August - Lead Data Scientist at Bitly

2014 June - Lead Data Scientist at Spring

bitly

Real-Time Media Map

1. Select a media type

Newspapers

2. Select your view

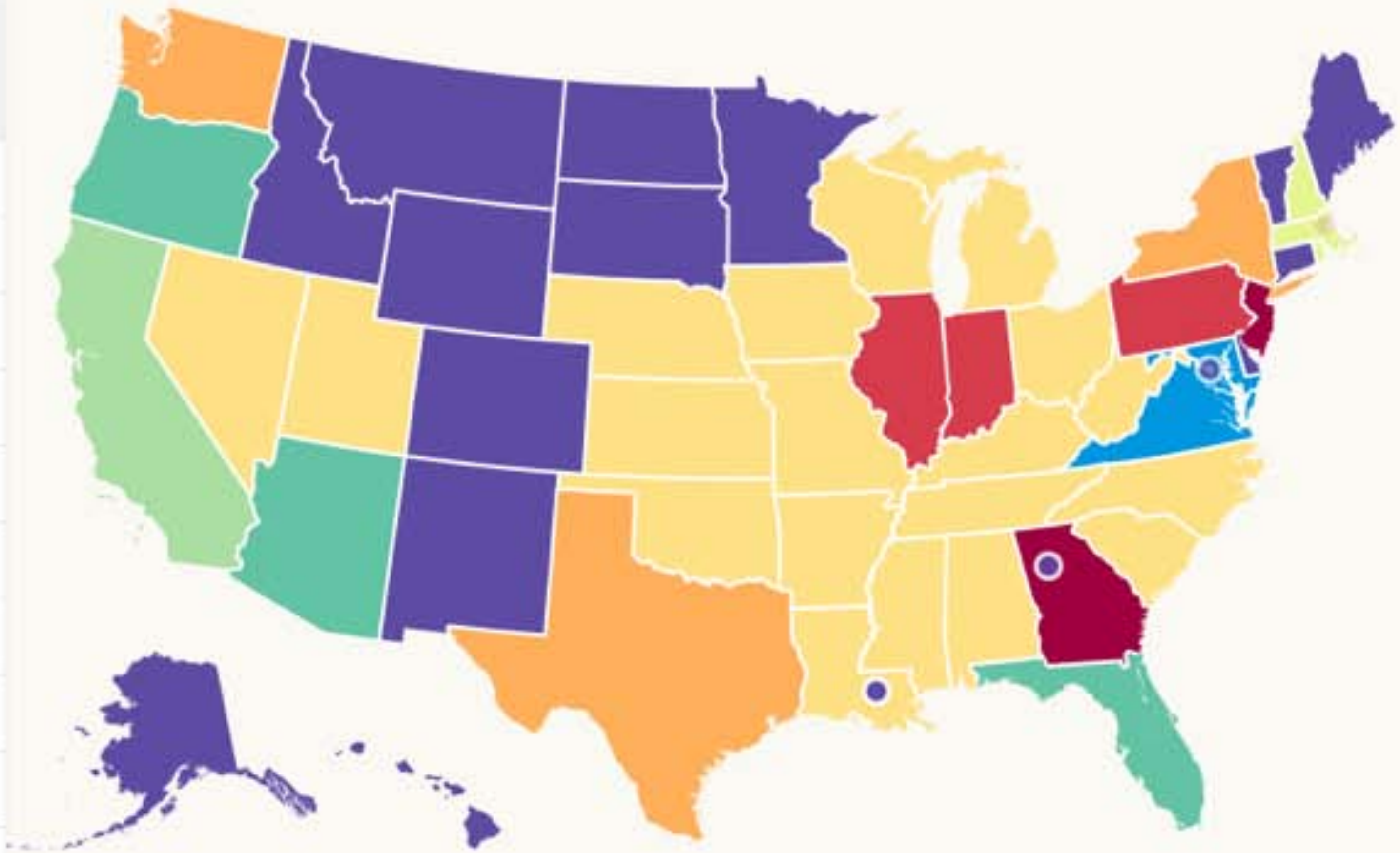
Real-time traffic

Winners by state

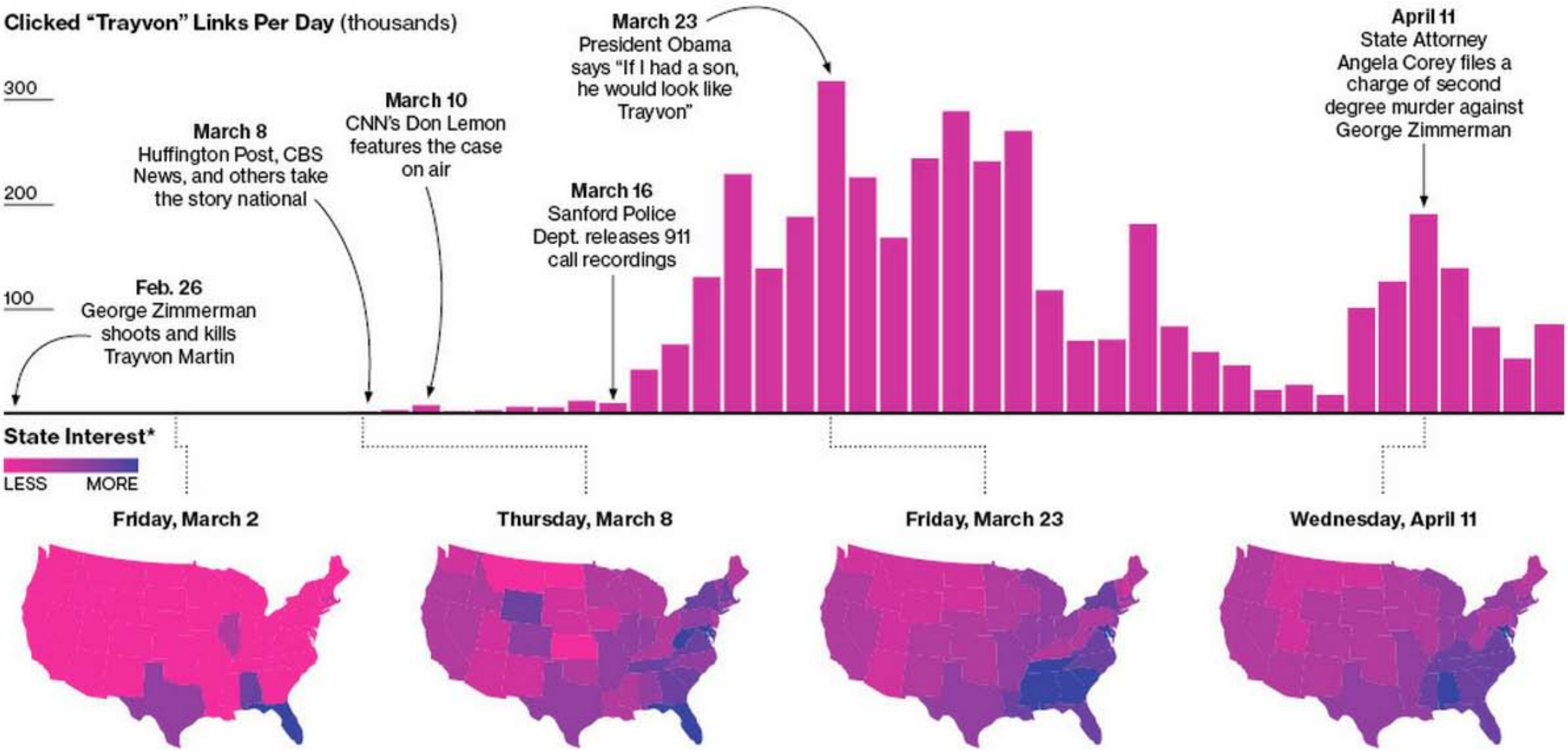
Both

Media Properties Legend
(click to view by property)

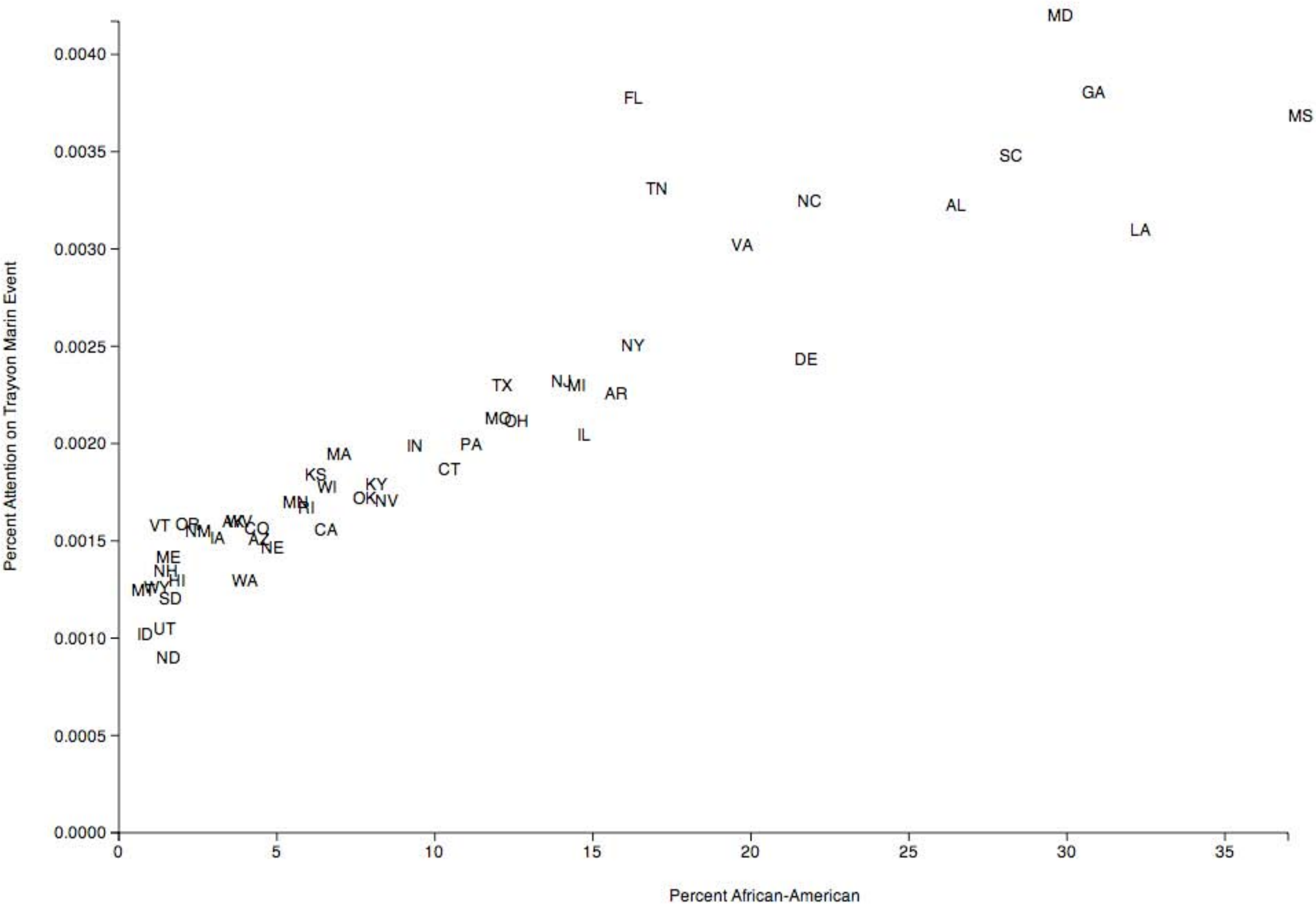
- The New York Times
- Washington Post
- Wall Street Journal
- Chicago Tribune
- Los Angeles Times
- SF Gate
- Boston Globe
- USA Today
- The Guardian
- San Jose Mercury



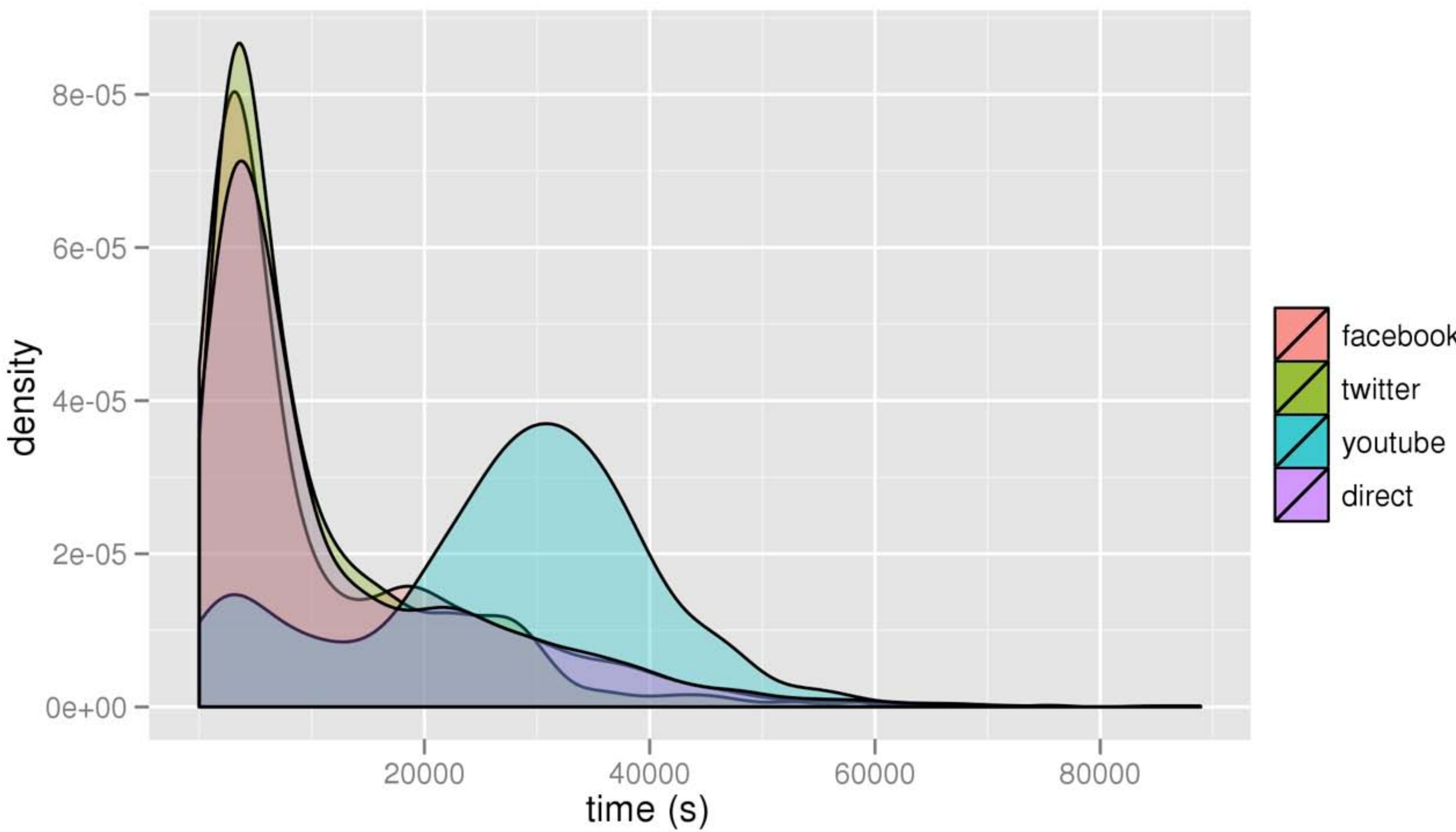
Interested in what Bitly can do with your link click data? Please reach out to us at community@bitly.com



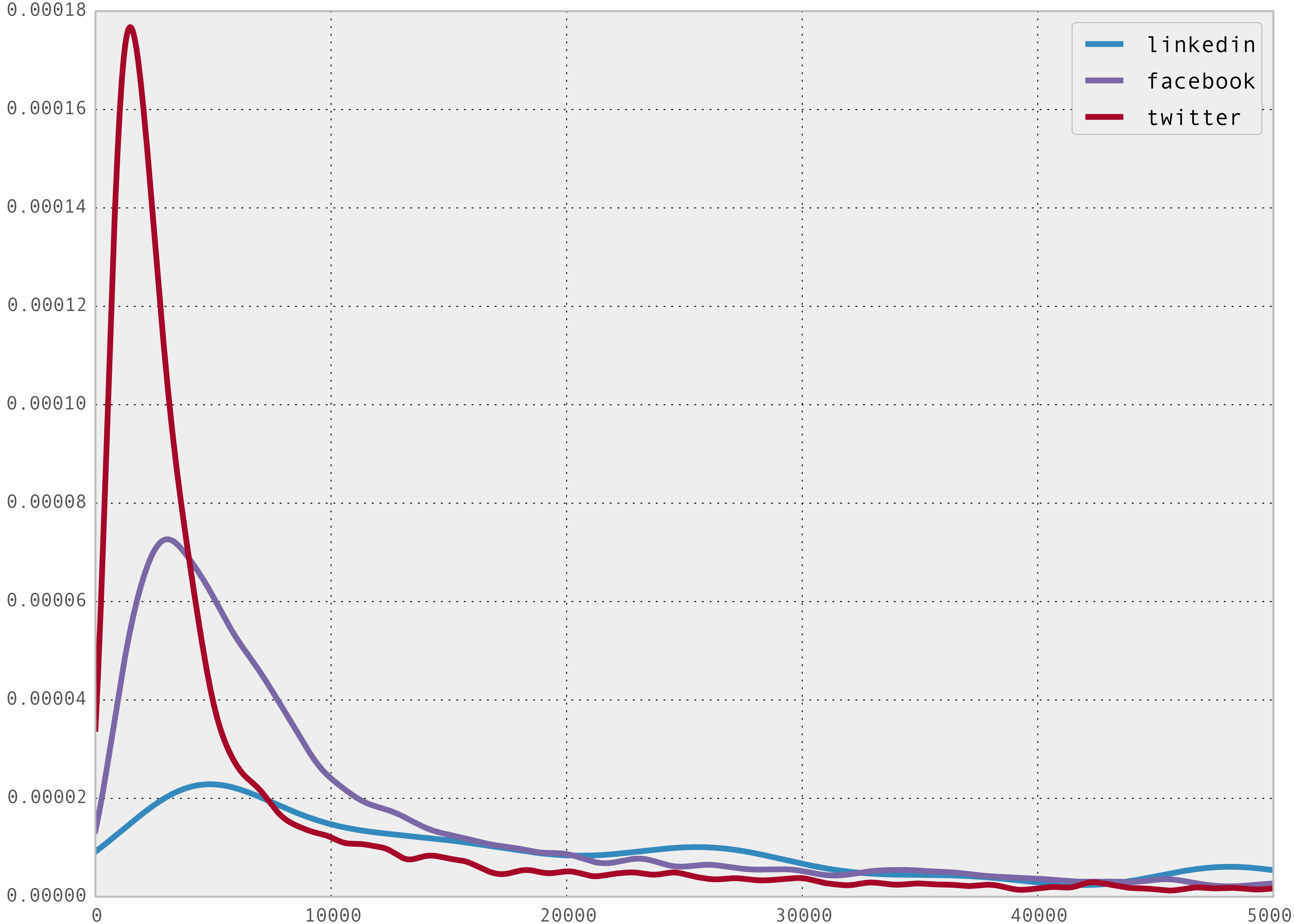
*STATE INTEREST IS THE PORTION OF CLICKED LINKS THAT RELATE TO TRAYVON MARTIN COMPARED WITH OTHER STATES IN ONE DAY; GRAPHIC BY BLOOMBERG BUSINESSWEEK; DATA: BITLY



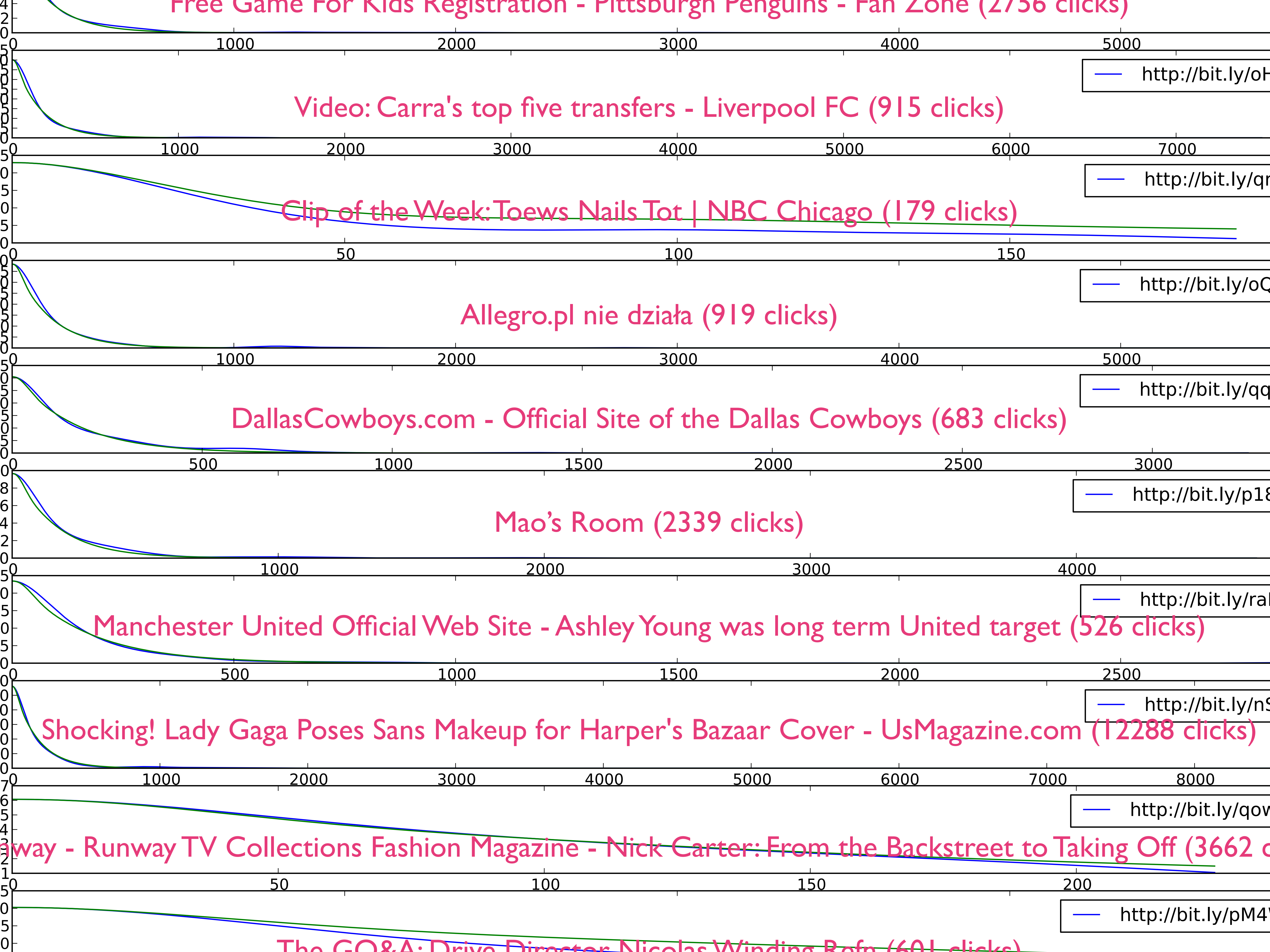
Percent Attention on Trayvon Marin Event
Percent African-American
0.0040
0.0035
0.0030
0.0025
0.0020
0.0015
0.0010
0.0005
0.0000
0
5
10
15
20
25
30
35
MD
GA
MS
FL
SC
TN
NC
AL
LA
VA
NY
DE
TX
NJ
MI
AR
MO
OH
IL
IN
PA
MA
CT
KS
WI
KY
OK
NV
MN
RI
VT
OR
NM
IA
CO
AZ
NE
CA
ME
NH
HI
WA
SD
ID
UT
ND

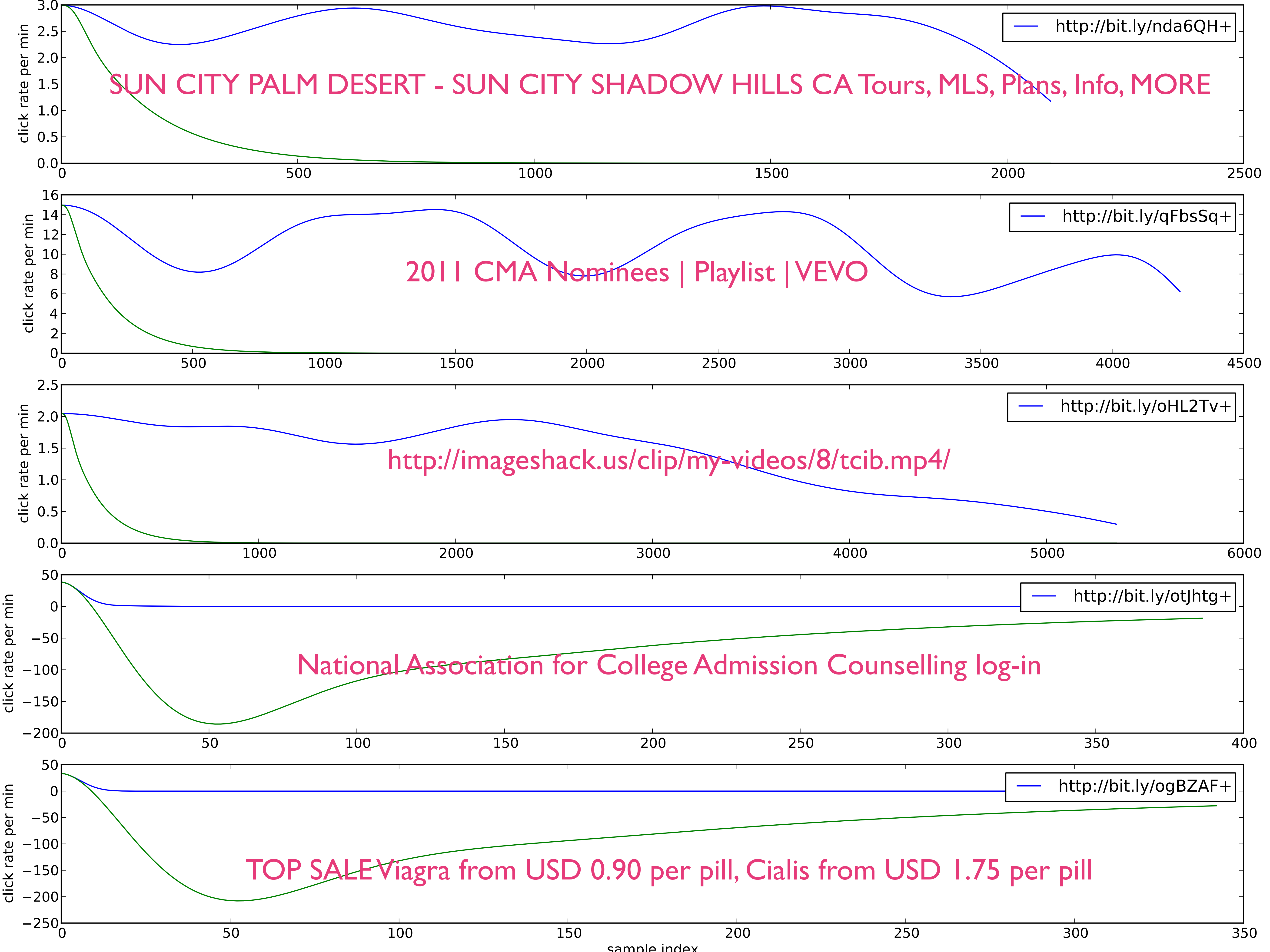


8e-05
6e-05
4e-05
2e-05
0e+00
density
20000
40000
60000
80000
time (s)
facebook
twitter
youtube
direct



0.00018
0.00016
0.00014
0.00012
0.00010
0.00008
0.00006
0.00004
0.00002
0.00000
0
10000
20000
30000
40000
50000
linkedin
facebook
twitter

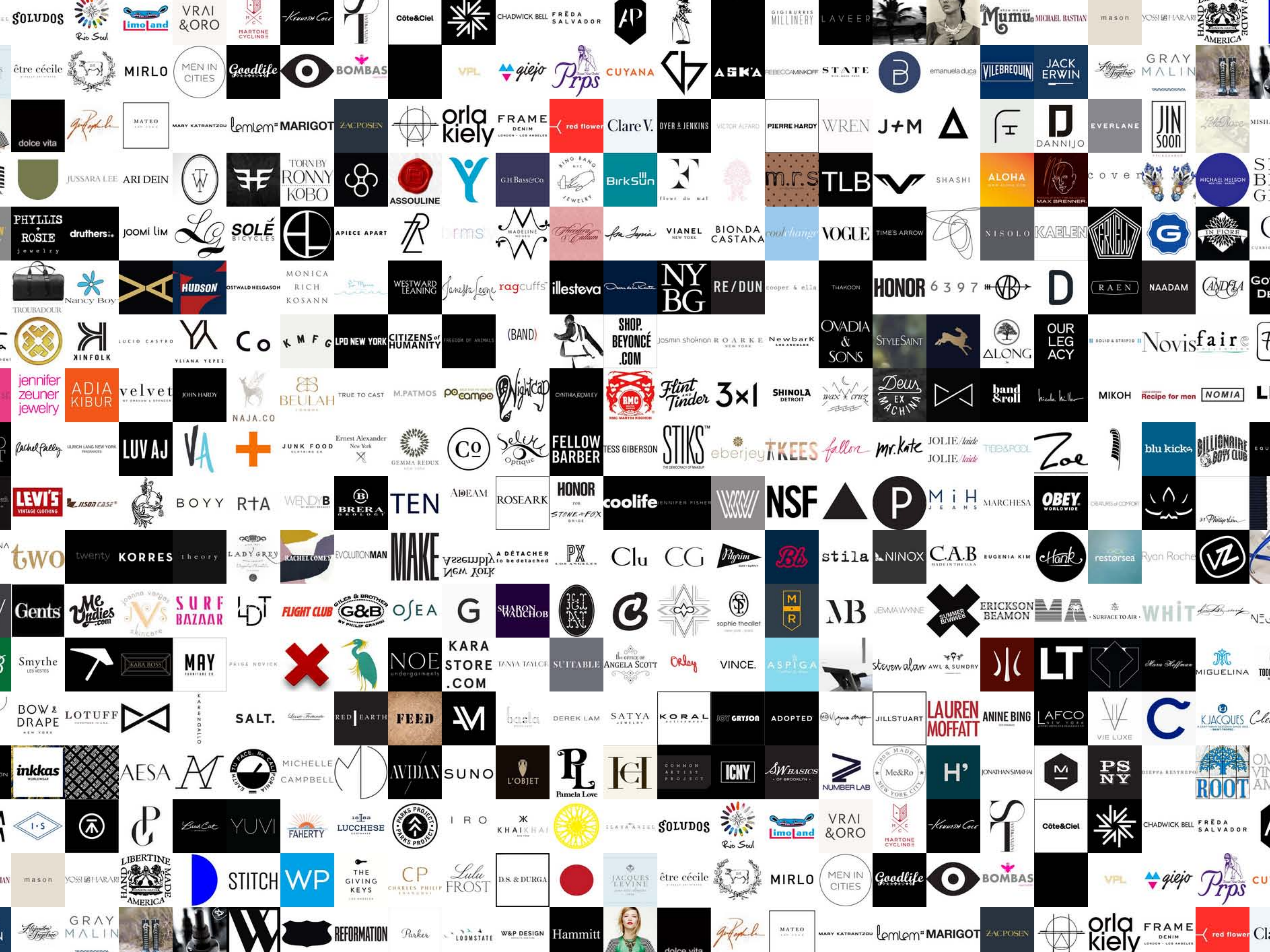

Free Game For Kids Registration - Pittsburgh Penguins - Fan Zone (2756 clicks)
http://bit.ly/oH
Video: Carra's top five transfers - Liverpool FC (915 clicks)
http://bit.ly/qr
Clip of the Week: Toews Nails Tot | NBC Chicago (179 clicks)
http://bit.ly/oQ
Allegro.pl nie działa (919 clicks)
http://bit.ly/qq
DallasCowboys.com - Official Site of the Dallas Cowboys (683 clicks)
http://bit.ly/p18
Mao's Room (2339 clicks)
http://bit.ly/ral
Manchester United Official Web Site - Ashley Young was long term United target (526 clicks)
http://bit.ly/n9
Shocking! Lady Gaga Poses Sans Makeup for Harper's Bazaar Cover - UsMagazine.com (12288 clicks)
http://bit.ly/qow
way - Runway TV Collections Fashion Magazine - Nick Carter: From the Backstreet to Taking Off (3662 c
http://bit.ly/pM4
The GQ&A: Drive Director Nicolas Winding Refn (601 clicks)


click rate per min
http://bit.ly/nda6QH+
SUN CITY PALM DESERT - SUN CITY SHADOW HILLS CA Tours, MLS, Plans, Info, MORE
http://bit.ly/qFbsSq+
2011 CMA Nominees | Playlist | VEVO
http://bit.ly/oHL2Tv+
http://imageshack.us/clip/my-videos/8/tcib.mp4/
http://bit.ly/otJhtg+
National Association for College Admission Counselling log-in
http://bit.ly/ogBZAF+
TOP SALE Viagra from USD 0.90 per pill, Cialis from USD 1.75 per pill
sample index


SPRING
EST 2013 NYC

Big-Data
to
No-Data


BELL
4:21 PM
SPRING
EST 2013 NYC


SOLUDOS
Rio Soul
Limoland
VRAI &ORO
MARTONE CYCLING
Côte&Ciel
CHADWICK BELL
FRĒDA SALVADOR
GIGI BURRIS MILLINERY
LAVEER
MICHAEL BASTIAN
mason
YOSSI HARARI
HAND MADE AMERICA
être cécile
MIRLO
MEN IN CITIES
Goodlife
BOMBAS
VPL
giejo
Prps
CUYANA
ASKA
REBECCA MINKOFF
STATE
emanuela duca
VILEBREQUIN
JACK ERWIN
GRAY MALIN
dolce vita
MATEO
MARY KATRANTZOU
lemlem
MARIGOT
ZAC POSEN
orla kiely
FRAME DENIM
red flower
Clare V.
DYER & JENKINS
VICTOR ALFARO
PIERRE HARDY
WREN
J+M
DANNIJO
EVERLANE
JIN SOON
MISHA
JUSSARA LEE
ARI DEIN
TORN BY RONNY KOBO
ASSOULINE
G.H.Bass&Co.
BirkSun
fleur du mal
m.r.s
TLB
SHASHI
ALOHA
MAX BRENNER
cover
MICHAEL NELSON
PHYLLIS + ROSIE jewelry
druthers
JOOMI LIM
SOLÉ BICYCLES
APIECE APART
MADELINE WEINRIB
VIANEL
BIONDA CASTANA
coolchange
VOGUE
TIMES ARROW
NISOLO
KAELEN
IN FIORE
TROUBADOUR
Nancy Boy
HUDSON
OSTWALD HELGASON
MONICA RICH KOSANN
WESTWARD LEANING
ragcuffs
illesteva
NY BG
RE/DUN
cooper & ella
THAKOON
HONOR
6397
RAEN
NAADAM
KINFOLK
LUCIO CASTRO
YLIANA YEPEZ
Co
KMFG
LPD NEW YORK
CITIZENS of HUMANITY
FREEDOM OF ANIMALS
(BAND)
SHOP. BEYONCÉ .COM
Josmin shokrion
ROARKE NEW YORK
Newbark
OVADIA & SONS
STYLESAINT
ALONG
OUR LEGACY
SOLID & STRIPED
Novis
faire
jennifer zeuner jewelry
ADIA KIBUR
velvet
JOHN HARDY
NAJA.CO
BEULAH LONDON
TRUE TO CAST
M.PATMOS
pocampo
Nightcap
CYNTHIA ROWLEY
RMC MARTIN ROSE
Flint and Tinder
3x1
SHINOLA DETROIT
Deus Ex Machina
band of outsiders
nicole miller
MIKOH
Recipe for men
NOMIA
Rachel Pally
ULRICH LANG NEW YORK
LUV AJ
JUNK FOOD
Ernest Alexander New York
GEMMA REDUX
Co
Selima Optique
FELLOW BARBER
TESS GIBERSON
STIKS
eberjey
TKEES
fallon
mr. kate
JOLIE laide
TIDE&POOL
Zoe
blu kicks
BILLIONAIRE BOYS CLUB
LEVI'S VINTAGE CLOTHING
BOYY
R+A
WENDY B
BRERA OROLOGI
TEN
ADEAM
ROSEARK
HONOR
STONE FOX BRIDE
coolife
JENNIFER FISHER
NSF
P
MiH JEANS
MARCHESA
OBEY WORLDWIDE
CREATURES of COMFORT
Phillip Lim
two
twenty
KORRES
theory
LADY GREY
RACHEL COMEY
EVOLUTIONMAN
MAKE
Assembly New York
A DÉTACHER
PX LOS ANGELES
Clu
CG
Pilgrim
Bb
stila
NINOX
C.A.B
EUGENIA KIM
Hank
restorsea
Ryan Roche
Gents
Me Undies .com
SURF BAZAAR
FLIGHT CLUB
G&B
OSEA
G
SHARON WAUCHOB
sophie theallet
MR
MB
JEMMA WYNNE
SUMMER BRAWNER
ERICKSON BEAMON
SURFACE TO AIR
WHIT
Smythe
KARA ROSS
MAY FURNITURE CO.
PAIGE NOVICK
NOE undergarments
KARA STORE .COM
TANYA TAYLOR
SUITABLE
The Office of Angela Scott
Orley
VINCE.
ASPIGA
steven alan
AWL & SUNDRY
LT
Mara Hoffman
MIGUELINA
BOW & DRAPE NEW YORK
LOTUFF
KARENZALLO
SALT.
RED EARTH
FEED
derek lam
SATYA
KORAL
JOY GRYSON
ADOPTED
JILLSTUART
LAUREN MOFFATT
ANINE BING
LAFCO
VIE LUXE
C
KJACQUES
inkkas
AESA
MICHELLE CAMPBELL
AVIDAN
SUNO
L'OBJET
Pamela Love
COMMON ARTIST PROJECT
ICNY
SW BASICS OF BROOKLYN
NUMBER LAB
Me&Ro
H'
JONATHAN SIMKHAI
PS NY
PIEPPA RESTREPO
ROOT
I-S
YUVI
FAHERTY
LUCCHESE
PARKS PROJECT
IRO
KHAI KHAI
SOLUDOS
Rio Soul
Limoland
VRAI &ORO
MARTONE CYCLING
Côte&Ciel
CHADWICK BELL
FRĒDA SALVADOR
mason
YOSSI HARARI
HAND MADE AMERICA
STITCH
WP
THE GIVING KEYS
CP CHARLES PHILIP
Lulu FROST
D.S. & DURGA
JACQUES LEVINE
être cécile
MIRLO
MEN IN CITIES
Goodlife
BOMBAS
VPL
giejo
Prps
GRAY MALIN
REFORMATION
Parker
LOOMSTATE
W&P DESIGN
Hammitt
dolce vita
MATEO
MARY KATRANTZOU
lemlem
MARIGOT
ZAC POSEN
orla kiely
FRAME DENIM
red flower


houzz
UBER
airbnb
STARBUCKS
SEPHORA
Jackthreads
Hotel Tonight
DISNEY STORE
GROUPON
Instacart
Eventbrite
Levi's STADIUM
TICKETS.COM
FANCY
Chairish
SPRING
EST 2013 NYC
Panera BREAD
STAPLES
StubHub!
OpenTable

Three Areas of Data Science

Data Engineering

Brian David Eoff @bde · Jun 11

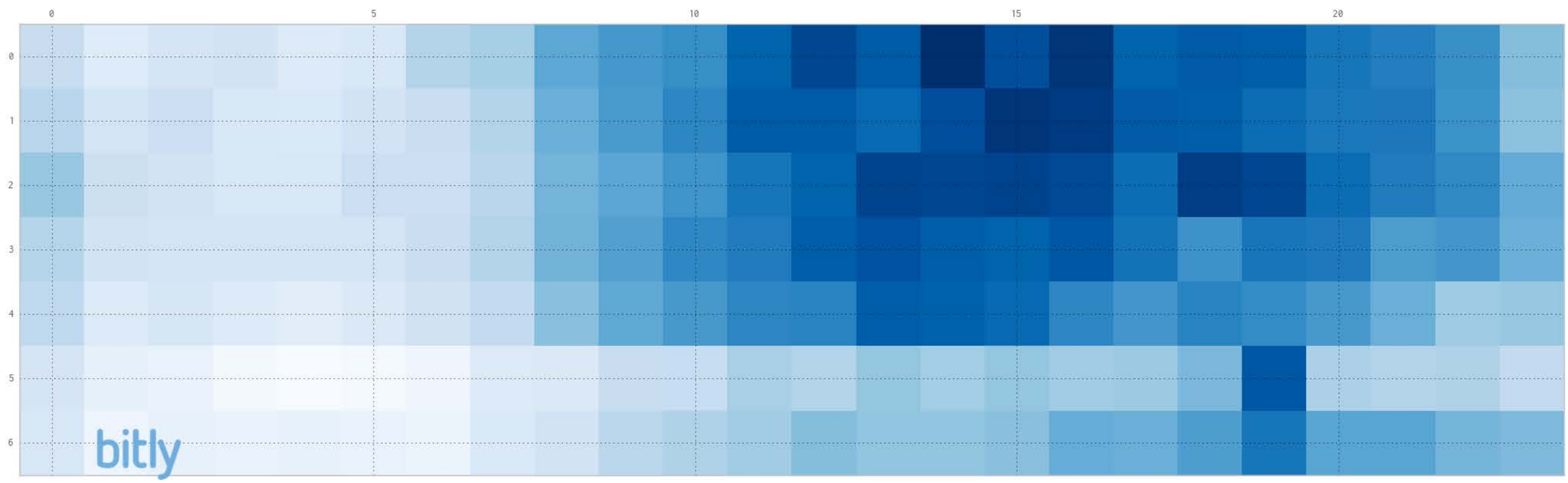
Not so much a data scientist as a data plumber. Conceiving of the flow, pipes and stores necessary to accomplish our goals.

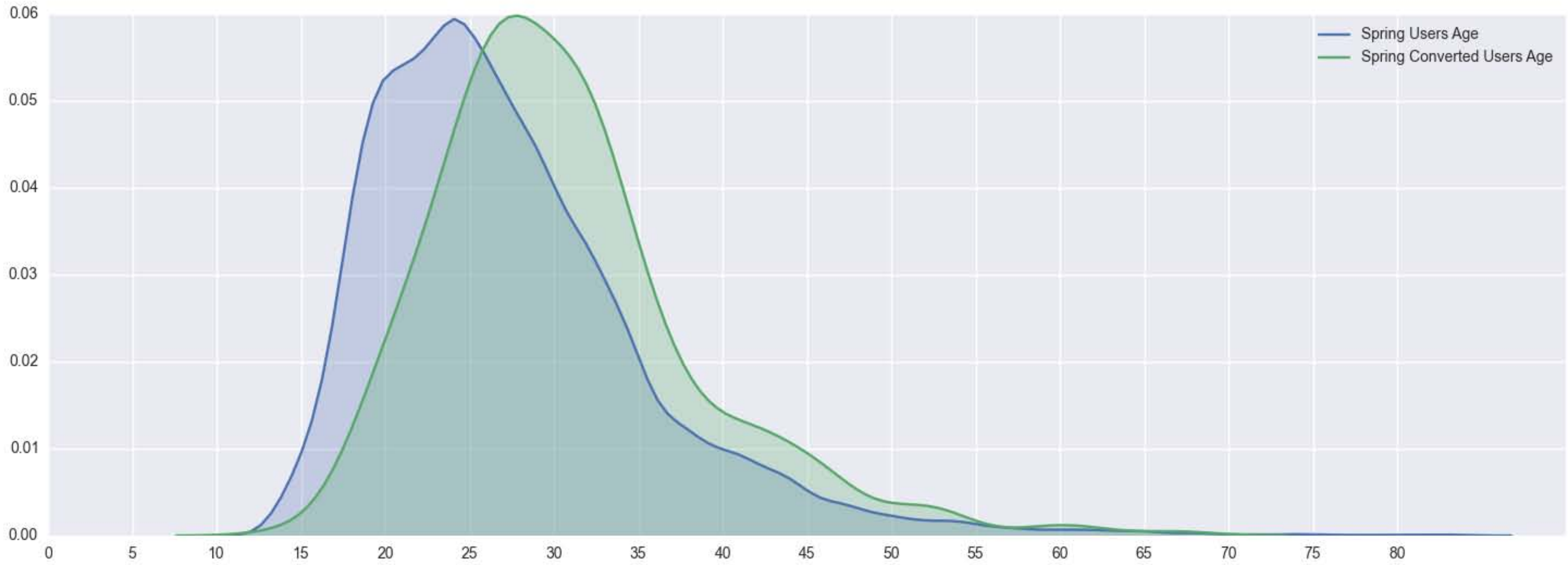
More than…

hadoop

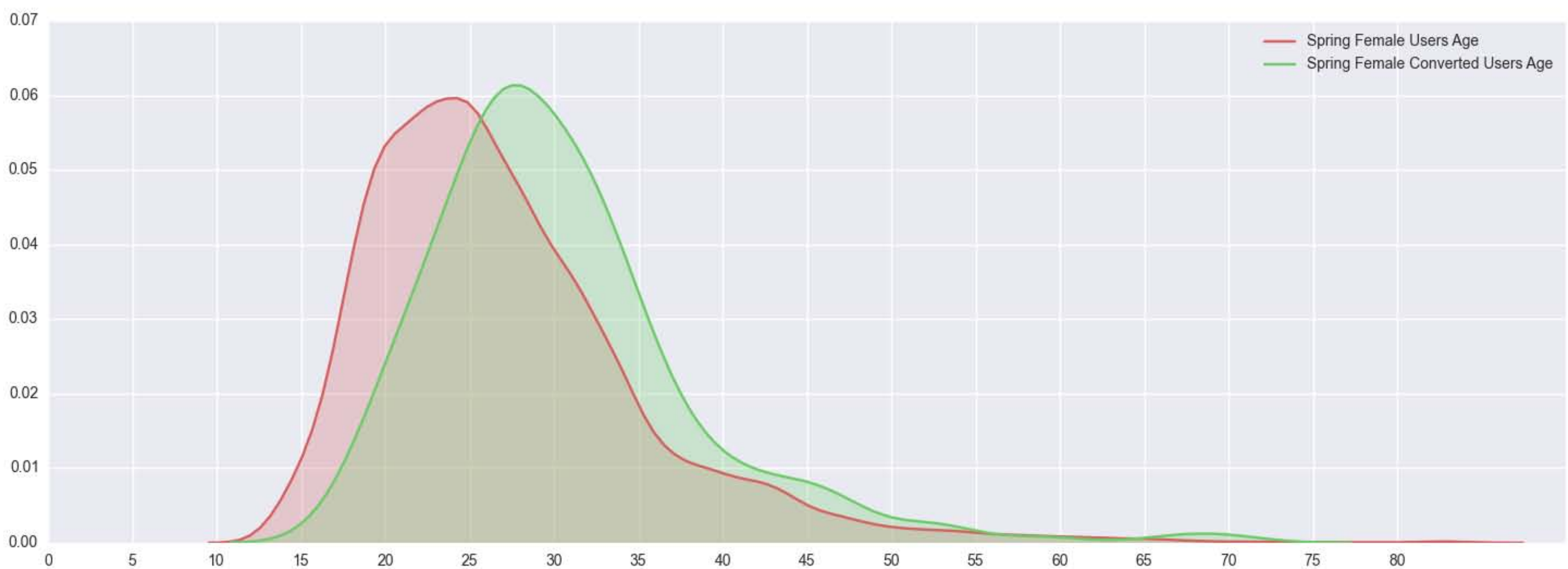
Everything in UTC

Data Analysis

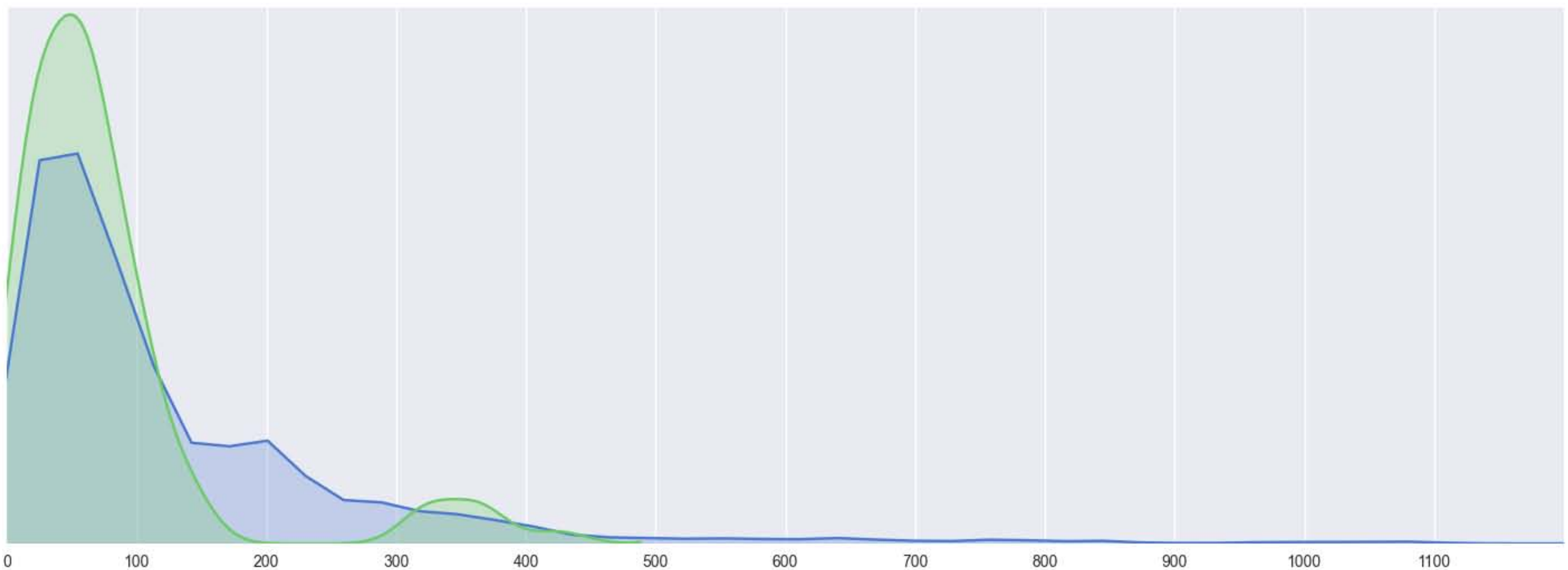

0
5
10
15
20
0
1
2
3
4
5
6
bitly



Spring Users Age
Spring Converted Users Age
0.06
0.05
0.04
0.03
0.02
0.01
0.00
0 5 10 15 20 25 30 35 40 45 50 55 60 65 70 75 80



Spring Female Users Age
Spring Female Converted Users Age
0.07
0.06
0.05
0.04
0.03
0.02
0.01
0.00
0 5 10 15 20 25 30 35 40 45 50 55 60 65 70 75 80



0
100
200
300
400
500
600
700
800
900
1000
1100

Reproducibility,
Automation,
Version Control

Data Modeling

Discovery

Collaborative Filtering for Implicit Feedback Datasets

Yifan Hu
AT&T Labs – Research
Florham Park, NJ 07932

Yehuda Koren*
Yahoo! Research
Haifa 31905, Israel

Chris Volinsky
AT&T Labs – Research
Florham Park, NJ 07932

Abstract

A common task of recommender systems is to improve customer experience through personalized recommendations based on prior implicit feedback. *These systems passively track different sorts of user behavior, such as purchase history, watching habits and browsing activity, in order to model user preferences. Unlike the much more extensively researched* explicit feedback*, we do not have any direct input from the users regarding their preferences. In particular, we lack substantial evidence on which products consumer dislike. In this work we identify unique properties of implicit feedback datasets. We propose treating the data as indication of positive and negative preference associated with vastly varying confidence levels. This leads to a factor model which is especially tailored for implicit feedback recommenders. We also suggest a scalable optimization procedure, which scales linearly with the data size. The algorithm is used successfully within a recommender system for television shows. It compares favorably with well tuned implementations of other known methods. In addition, we offer a novel way to give explanations to recommendations given by this factor model.*

tent based approach creates a profile for each user or product to characterize its nature. As an example, a movie profile could include attributes regarding its genre, the participating actors, its box office popularity, etc. User profiles might include demographic information or answers to a suitable questionnaire. The resulting profiles allow programs to associate users with matching products. However, content based strategies require gathering external information that might not be available or easy to collect.

An alternative strategy, our focus in this work, relies only on past user behavior without requiring the creation of explicit profiles. This approach is known as *Collaborative Filtering* (CF), a term coined by the developers of the first recommender system - Tapestry [8]. CF analyzes relationships between users and interdependencies among products, in order to identify new user-item associations. For example, some CF systems identify pairs of items that tend to be rated similarly or like-minded users with similar history of rating or purchasing to deduce unknown relationships between users and items. The only required information is the past behavior of users, which might be their previous transactions or the way they rate products. A major appeal of CF is that it is domain free, yet it can address aspects of the data that are often elusive and very difficult to profile using con-

Popularity

Tools

Caverlee’s Rule

How to become a data scientist?

CONFESSION

I never took a data science class.

What do I look for when I hire?

Questions?

Descriptive, Prescriptive, Predictive

Data Engineering, Data Analysis, Machine Learning/Model Building/Algorithms

ETL, or never underestimate the benefit of a good plumber.

AB Test. Multi-Armed Bandit Testing. Optimization. Retention

Bitly: 200 Million Events per Day
Spring: 0

Data Science (Inductive Reasoning), obtaining data, cleaning munging data, exploring data, modeling data interpreting data

Data Engineering

ETL

Redshift

Event Streams
Kafka, Kinesis, NSQ

Tools, Process, Recruiting, Communication, Ethics

Test- Driven Data Science

Learn from software development.

Version Control, Automated Testing

Analysis

Price Points

Conversion

Metric Aggregation

etc.

Matrix Factorization

Recommendation

Popularity