#DataScience

Russell Chen, Jin Rou New, Yanqiao Wang, Mao Zhou

Motivation

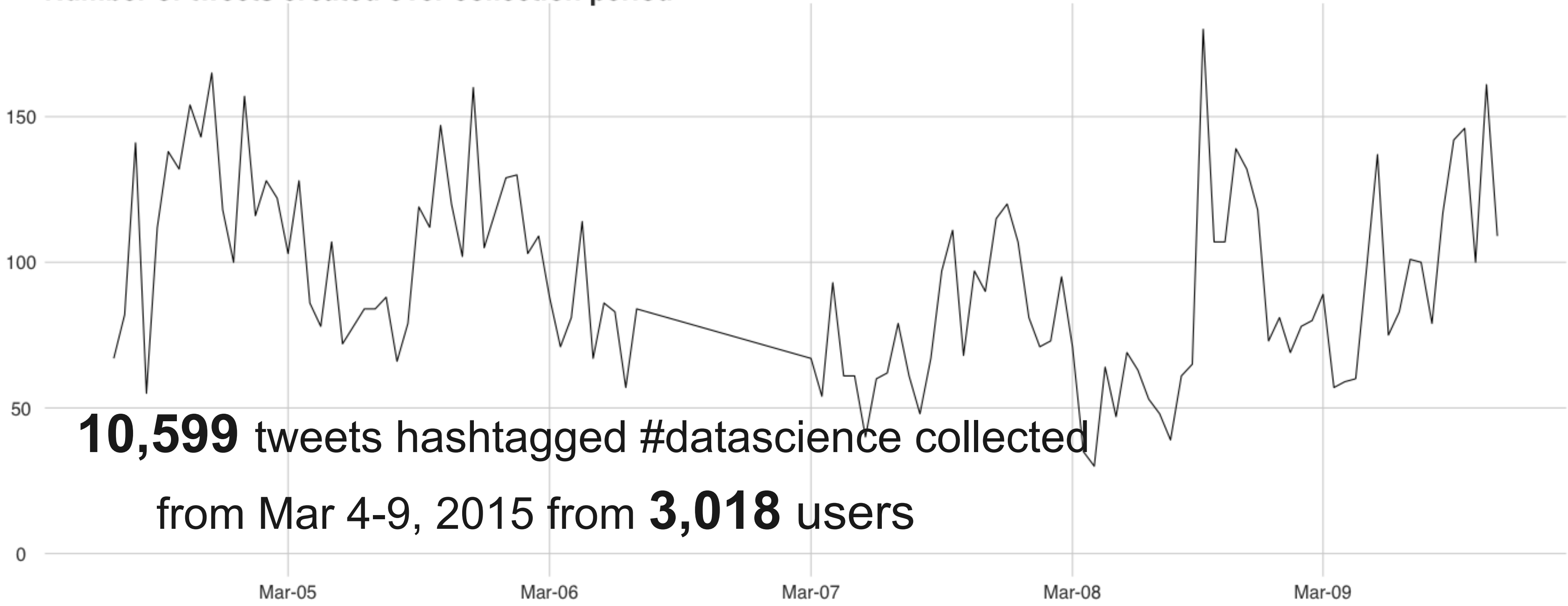
What is data science?

When did the conversation about data science start?

Who are the people who talk about data science?

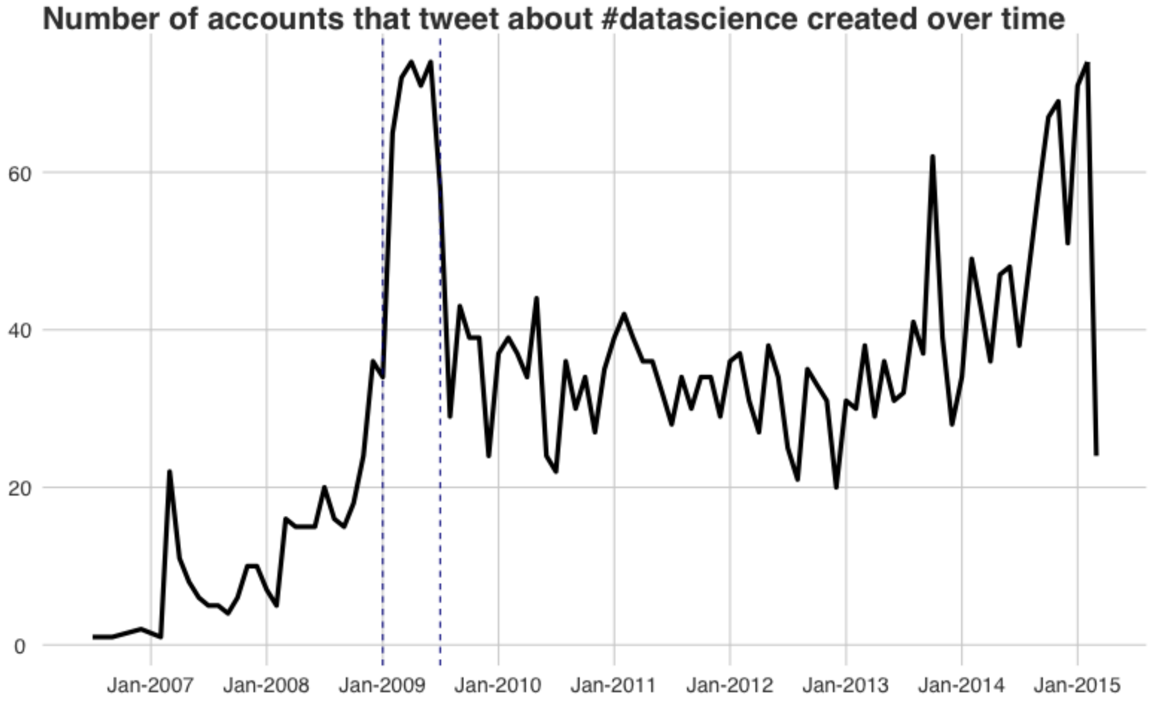
Data snapshot

Number of tweets created over collection period

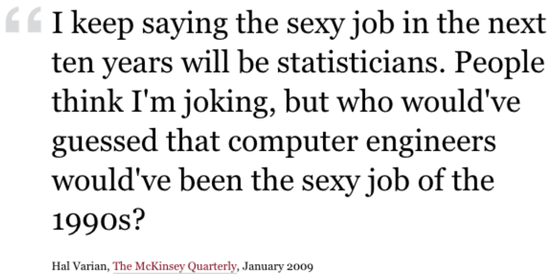


10,599 tweets hashtagged #datascience collected from Mar 4-9, 2015 from **3,018** users

When did #datascience become a thing?

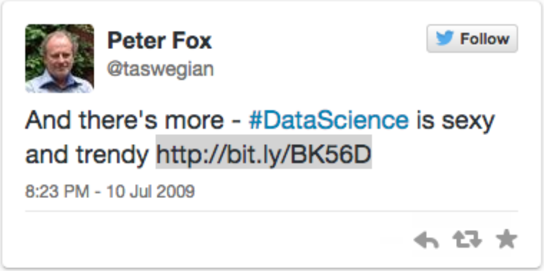


Number of accounts that tweet about #datascience created over time



> I keep saying the sexy job in the next ten years will be statisticians. People think I'm joking, but who would've guessed that computer engineers would've been the sexy job of the 1990s?

Hal Varian, The McKinsey Quarterly, January 2009



Peter Fox
@taswegian

And there's more - #DataScience is sexy and trendy http://bit.ly/BK56D

8:23 PM - 10 Jul 2009

What is #datascience?

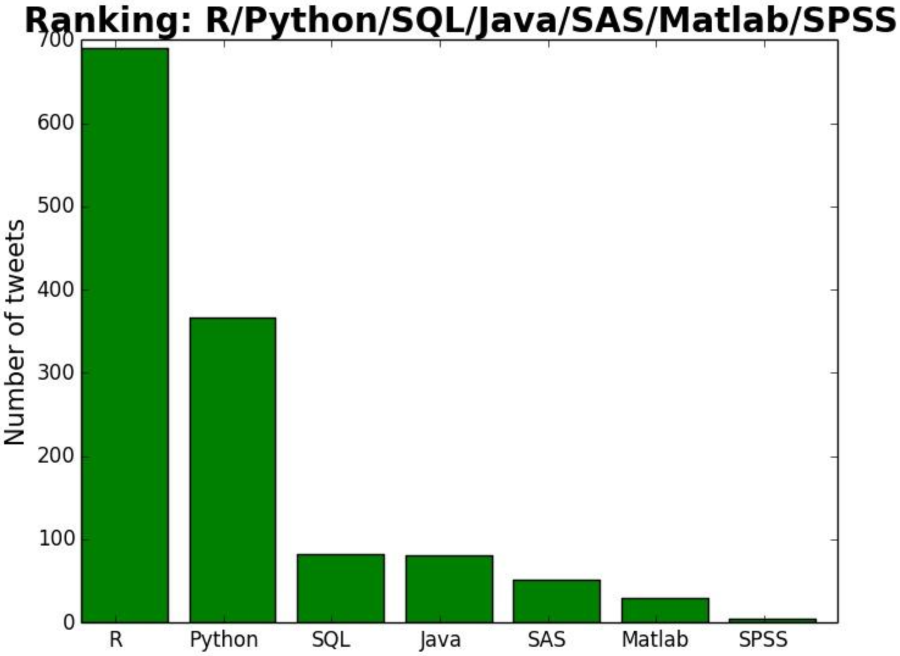
datascience
analytics
bigdata
machinelearning
statistics
datascientist
datamining
dataviz
hadoop
python
iot
rstats
internetofthings
predictiveanalytics
dataanalytics
datavisualization
deeplearning
datascientest
abdsc
healthcare
infographic
business
cloud
compsci
linkeddata
ai
opendata
marketing
bi
spark
stem
tech
java
education
software
crm
math
privacy
socialmedia
startup
ferguson
salesforce
mobile
artificialintelligence
nosql
businessintelligence
wearable
smartcity
javascript
mediterraneancoin
dataanalysis
m2m
apr360
mathchat
fintech
innovation
sql
infosec
opensource
ml
enterprise
devops
mapreduce
algorithm
ux
d3js
job
insight
analyst
technology
datascibowl

Word cloud for **hashtags**

learning
big
top
need
new
analytic
resource
using
machine
scientists
world
data
learn
will
make
business
future
follow
video
algorithm
article
mining
now
blog
skill
smart
work
time
tutorial
python
challenge
thing
set
best
change
better
guide
introduction
team
people
statistical
model
mean
way
map
trend
graph
full
use
job
power
thank
modern
look
analysis
tool
forums
awesome
women
interesting
deep
companies
step
professional
know
post
great
check
problem
infographic
open
good
first

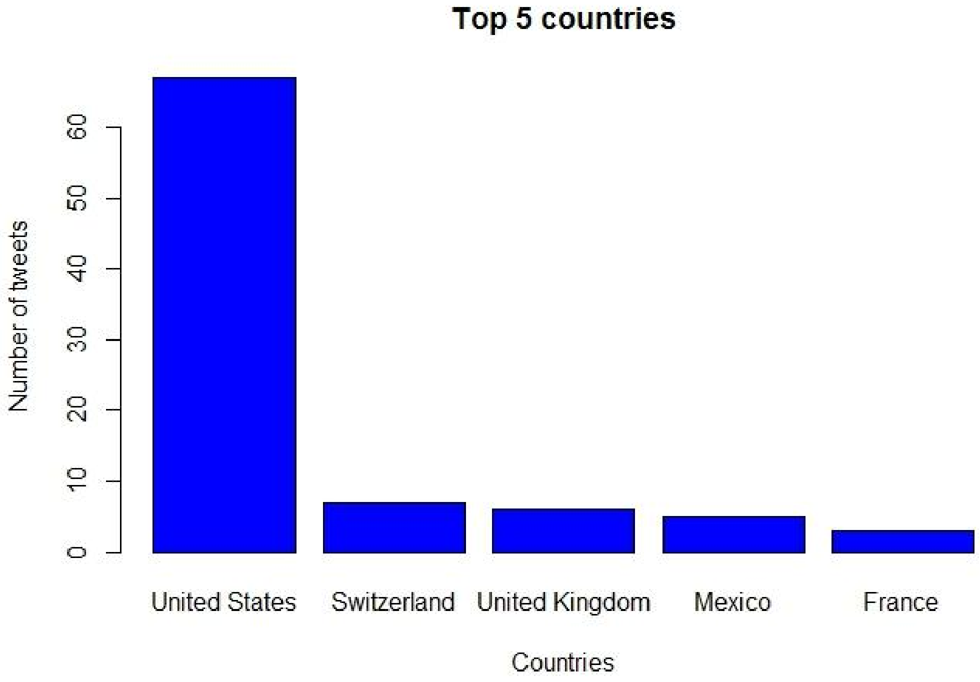
Word cloud for **tweets**

Comparison of programming languages



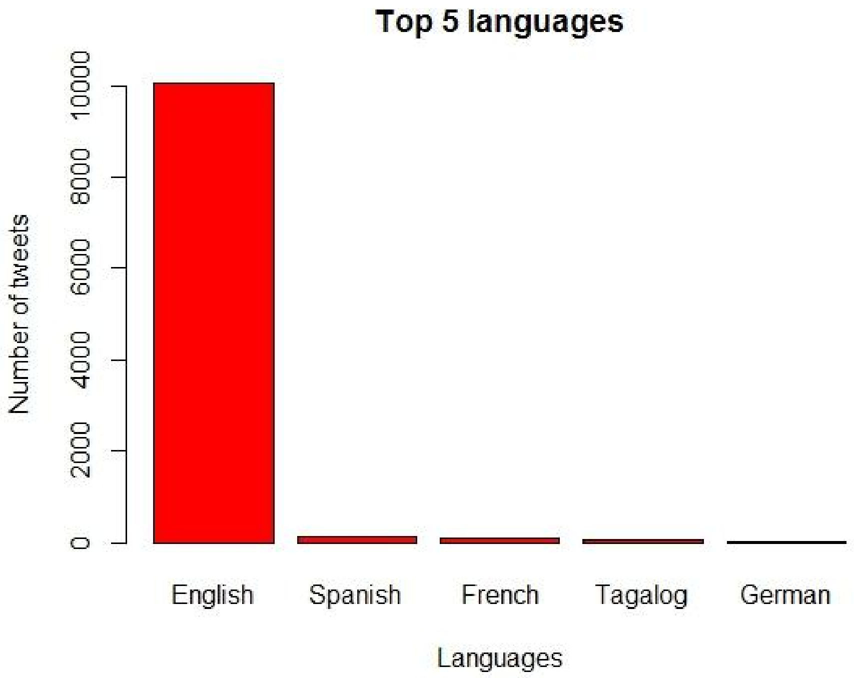
- Calculate the number of tweets for each programming language
- For the tweets that contain #datascience, R and Python seem to be dominant

Top 5 countries



- Top 5 countries from which the tweets were sent
- United States sent most of them

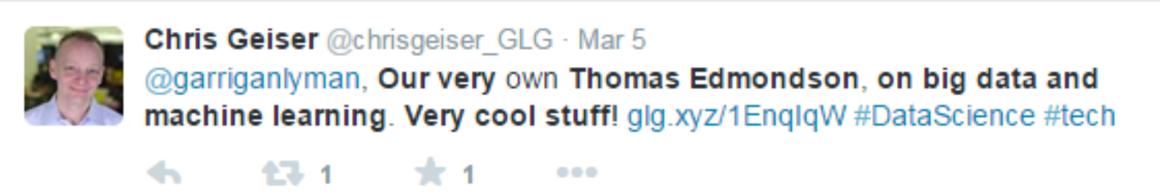
Top 5 languages



- Top 5 languages in which the tweets were written
- English is dominant

Sentiment analysis

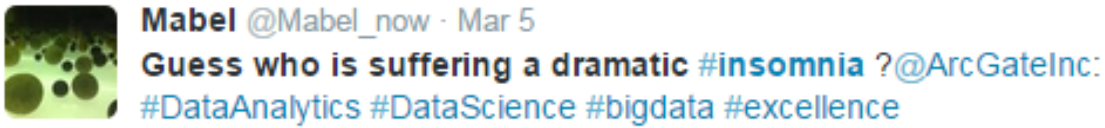
- We assume that each tweet has a **hidden sentiment** and that the words in a tweet are drawn from a **multinomial distribution** that depends only on its sentiment.
- Calculate the probability of each tweet being “Happy” or “Sad”.
- Results:
  - Happy: 67.78%. E.g.



Chris Geiser @chrisgeiser_GLG · Mar 5

@garriganlyman, **Our very own Thomas Edmondson, on big data and machine learning. Very cool stuff!** glg.xyz/1EnqlqW #DataScience #tech

  - Sad: 0.5%. E.g.



Mabel @Mabel_now · Mar 5

Guess who is suffering a dramatic #insomnia ?@ArcGateInc: #DataAnalytics #DataScience #bigdata #excellence

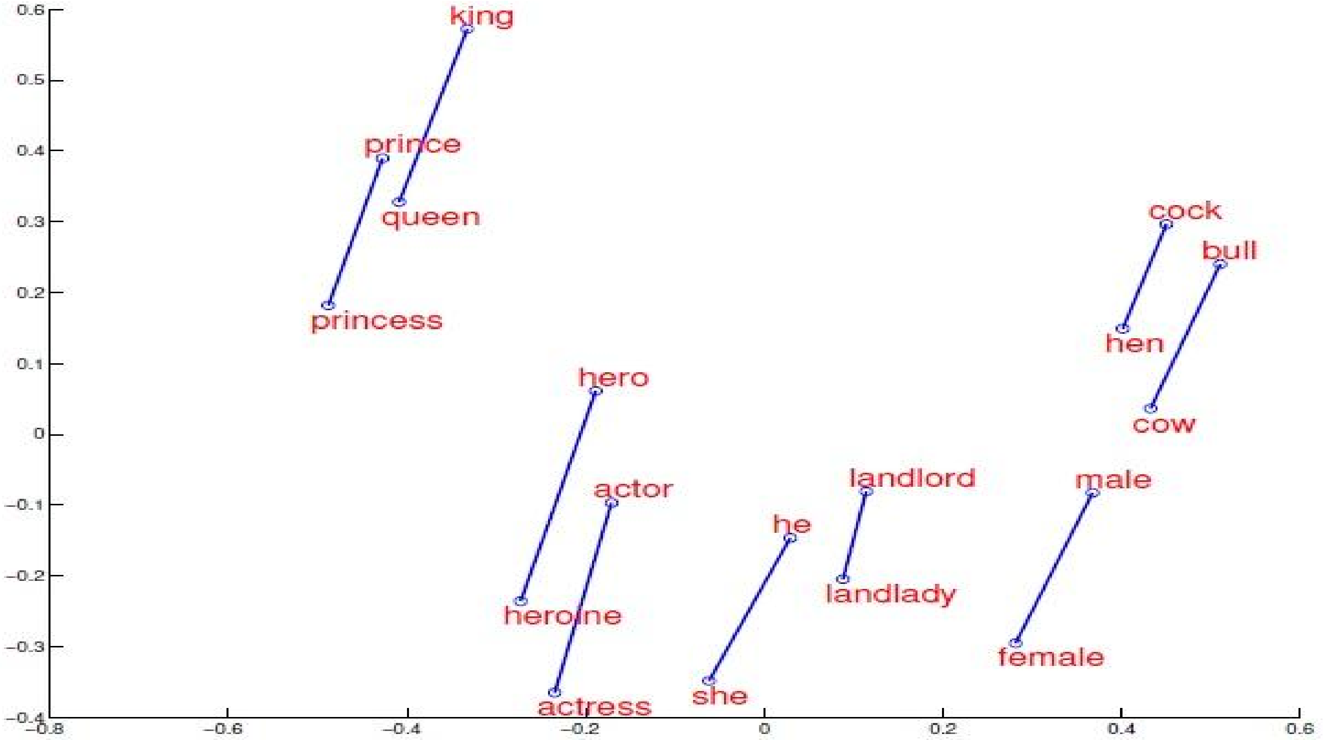
What are the topics and keywords in #datascience tweets?

Topic 1	Topic 2	Topic 3
machinelearn	analyt	statist
learn	bigdata	rstat
machin	hadoop	datasci
model	job	comput
tech	market	food

What are the topics and keywords in #datascience user profiles?

Topic 1	**Topic 2**	**Topic 3**
research	analyt	tech
phd	bigdata	entrepreneur
univers	datasci	startup
student	cloud	innov
interest	iot	consult

word2vec (Google 2013)



Training word2vec

Input:

[[‘this’, ‘is’, ‘my’, ‘first’, ‘sentence’],
  [‘and’, ‘now’, ‘for’, ‘my’, ‘second’, ‘sentence’]]

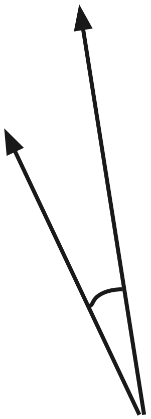
Output (for each word):

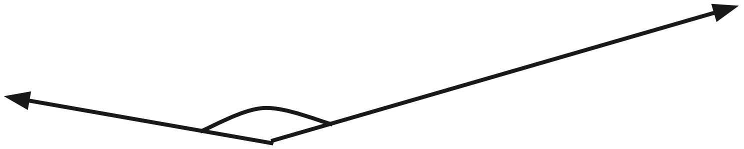
```
array([-0.00449447, -0.00310097,  0.02421786, ...], dtype=float32)
```

Word similarities

Cosine distance

$$\frac{a \circ b}{|a||b|}$$





‘stats’ + ‘analytics’ - ‘math’

‘datascience’

One of these is not like the others

‘datascience’

[✓ ‘python’]

‘machinelearning’

‘statistics’

Author influence for #datascience

- Easiest way to measure author influence is by number of followers
- Ways to increase your reach if you tweet about #datascience:
  - statuses_count: The total number of tweets and retweets a Twitter user has posted
  - friends_count: The number of Twitter friends the user has

Author influence for #datascience

- listed_count: The number of Twitter lists on which the author of a Tweet appears.
- word_count: If top ranked keywords could increase the authors influence; The frequency of top 20 hashtags are counted and summed.
- Other possible variables: location, language, age…
- Two approaches: Generalized Linear Model (Poisson regression) and Support Vector Machine

Author influence for #datascience (GLM)

- Adding certain keywords won’t raise your profile (P>0.5)
- Less influential you will be if you tweet more (negative coefficient)
- The best way to be influential is adding friends
- Not surprisingly, more influential people will have more people replying to their tweets

Author influence for #datascience (SVM)

- Simple two class classifier:
  - Class 1: 1-1000 followers
  - Class 2: > 1000 followers

Author influence for #datascience (Results)

Class/Prediction (# followers)	# statuses	# friends	# lists	# keywords
1~1000	10000	1000	100	1
881	10000	1000	100	1
>1000	67869	79747	981	2
124581	67869	79747	981	2

#ThankYou

Russell Chen, Jin Rou New, Yanqiao Wang, Mao Zhou