

[illegible]

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



Questions

When did the conversation about data science start?

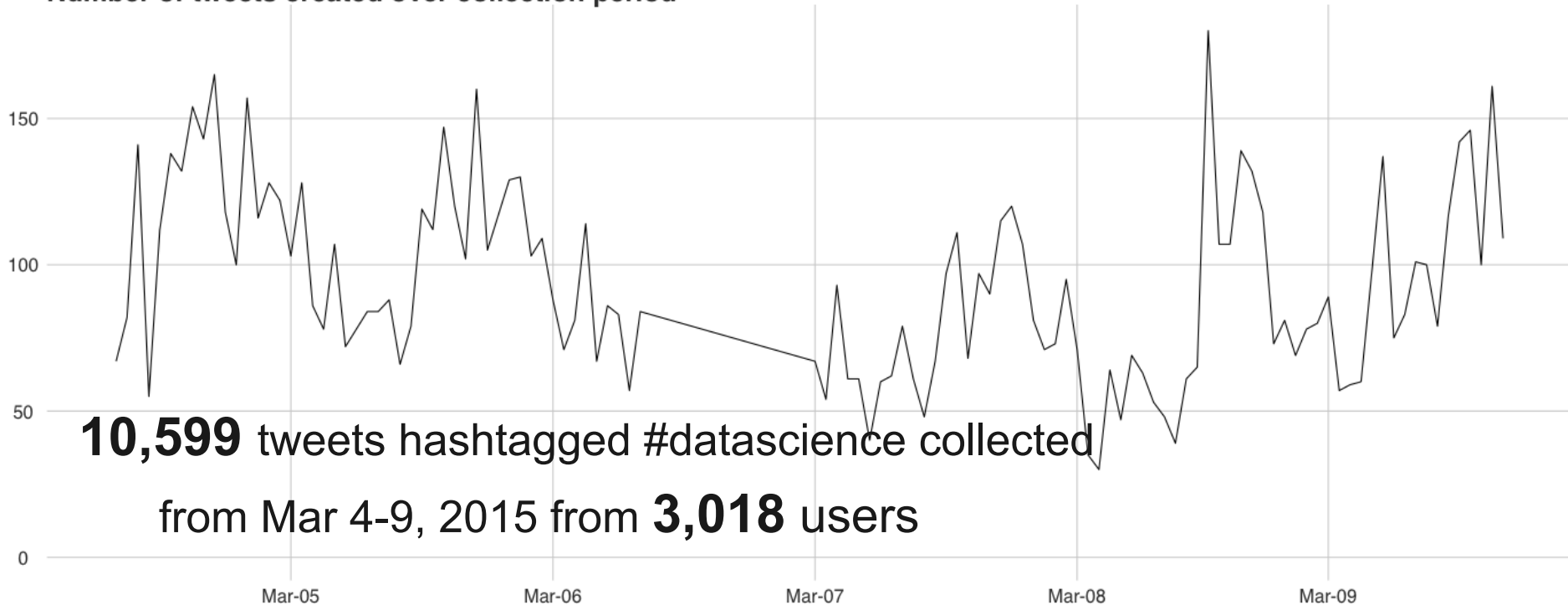
Who are those tweeting about data science?

What is data science?

How can you influence the data science conversation?

Data snapshot

Number of tweets created over collection period



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

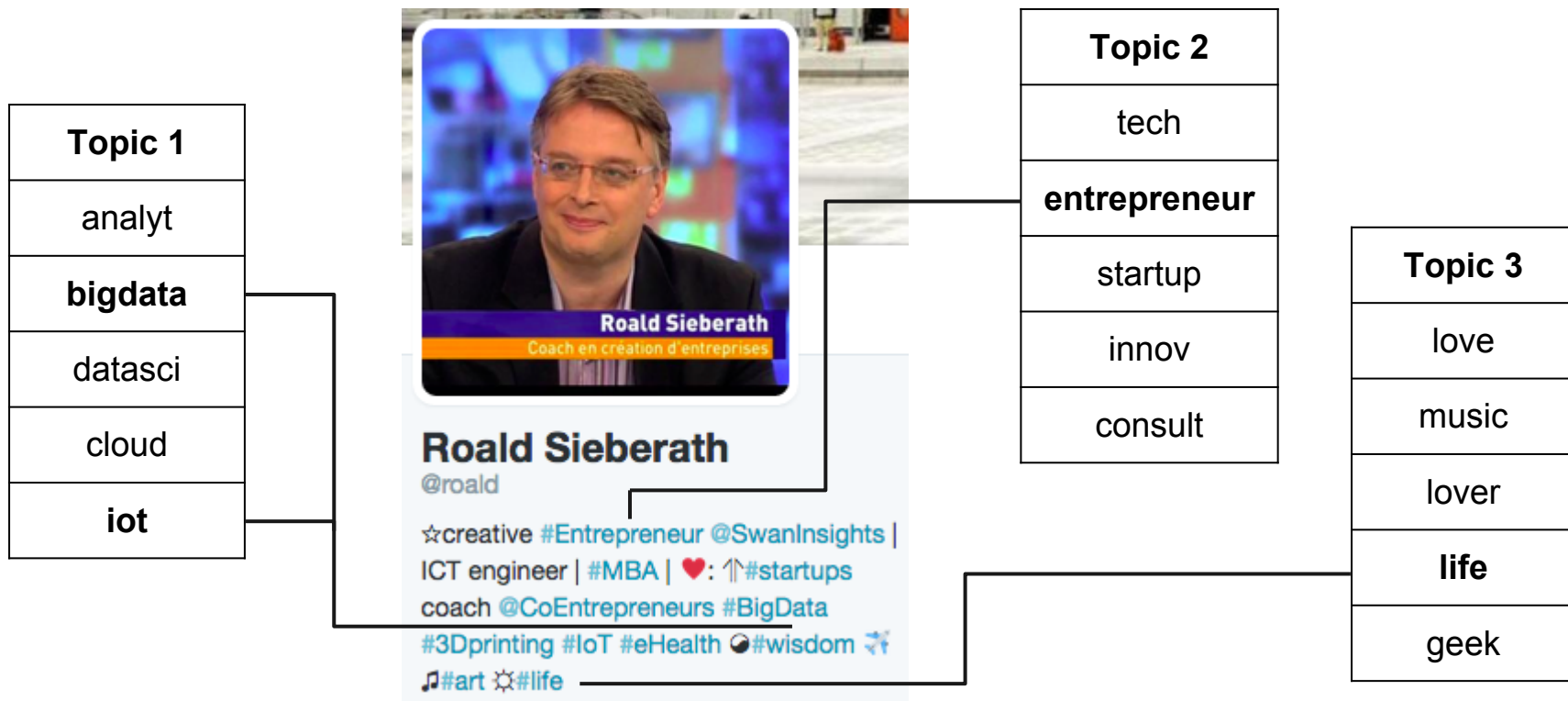
Follow

And there's more - [#DataScience](http://bit.ly/BK56D) is sexy and trendy <http://bit.ly/BK56D>

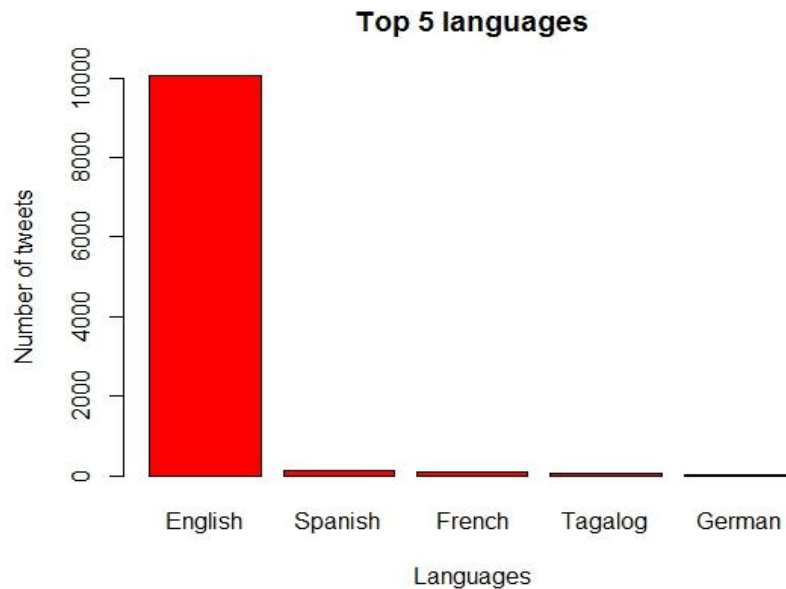
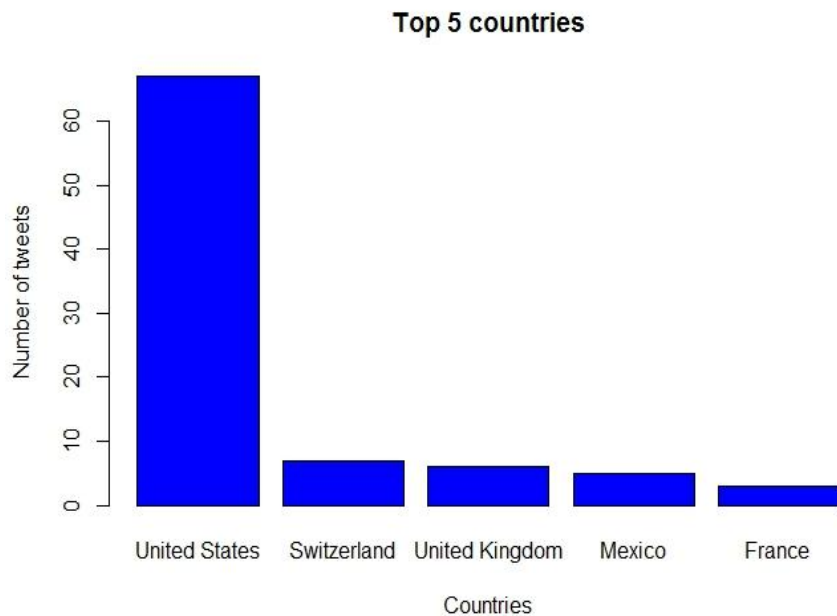
8:23 PM - 10 Jul 2009



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



What are the top countries and languages of #datascience users?



What is #datascience?



Word cloud for **hashtags**




Word cloud for tweets

What are the topics and keywords in #datascience tweets?

 **Slipstream Data** @SlipstreamZA · Mar 9
"What is [#Hadoop](#)? Explaining Big Data to The C-Suite" - ow.ly/K4JWT [#BigData](#) [#DataScience](#) [#BI](#)
🔄 5 🌟 2 ... [View summary](#)

Topic 1

bigdata, **datasci**,
hadoop, **analyt**, **job**

 **Lillian Pierson, PE** @BigDataGal · Mar 8
[#rstats](#) [#datascience](#) Some Intuition About the Theory of Statistical Learning
htl.li/2VEAP3 fb.me/2QXq3puAY
🔄 4 🌟 8 ...

Topic 2

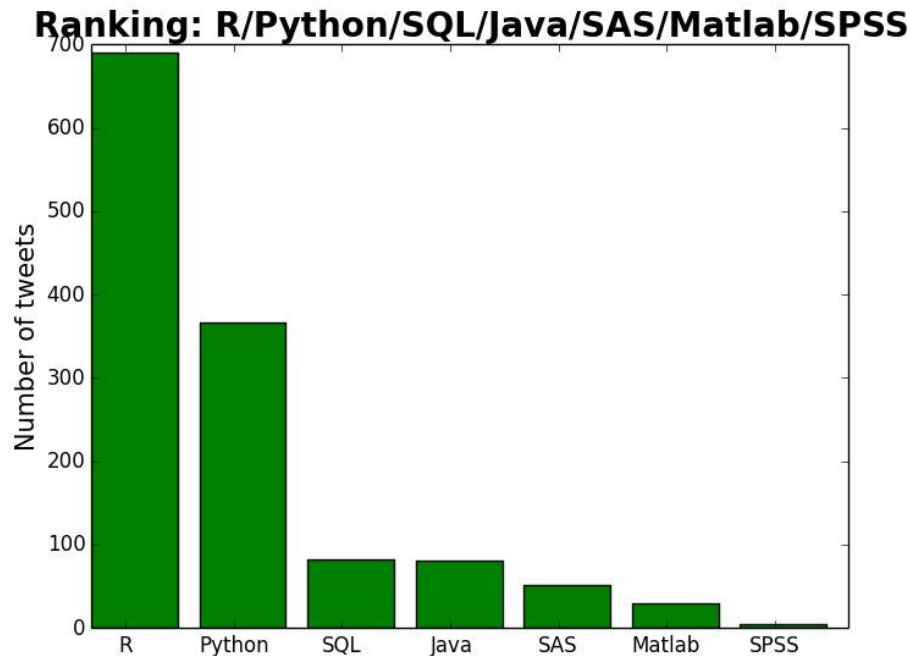
statist, **datasci**, **rstat**,
follow, now

 **Ekimetrics.** @ekimetrics · Feb 28
The full [#DataScientist](#) Skill Set via ekimetrics.com
[#DataScience](#) [#BigData](#) [#Dataviz](#)
[#MachinLearning](#) [#IoT](#) [#AI](#)

Topic 3

datasci, **datascientist**,
dataviz, **skill**, **set**

What are the top programming languages in #datascience?



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

Sentiment analysis

- Assumptions:
 - Each tweet has a **hidden sentiment**;
 - 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: 67.8%. 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



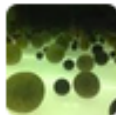
1



1



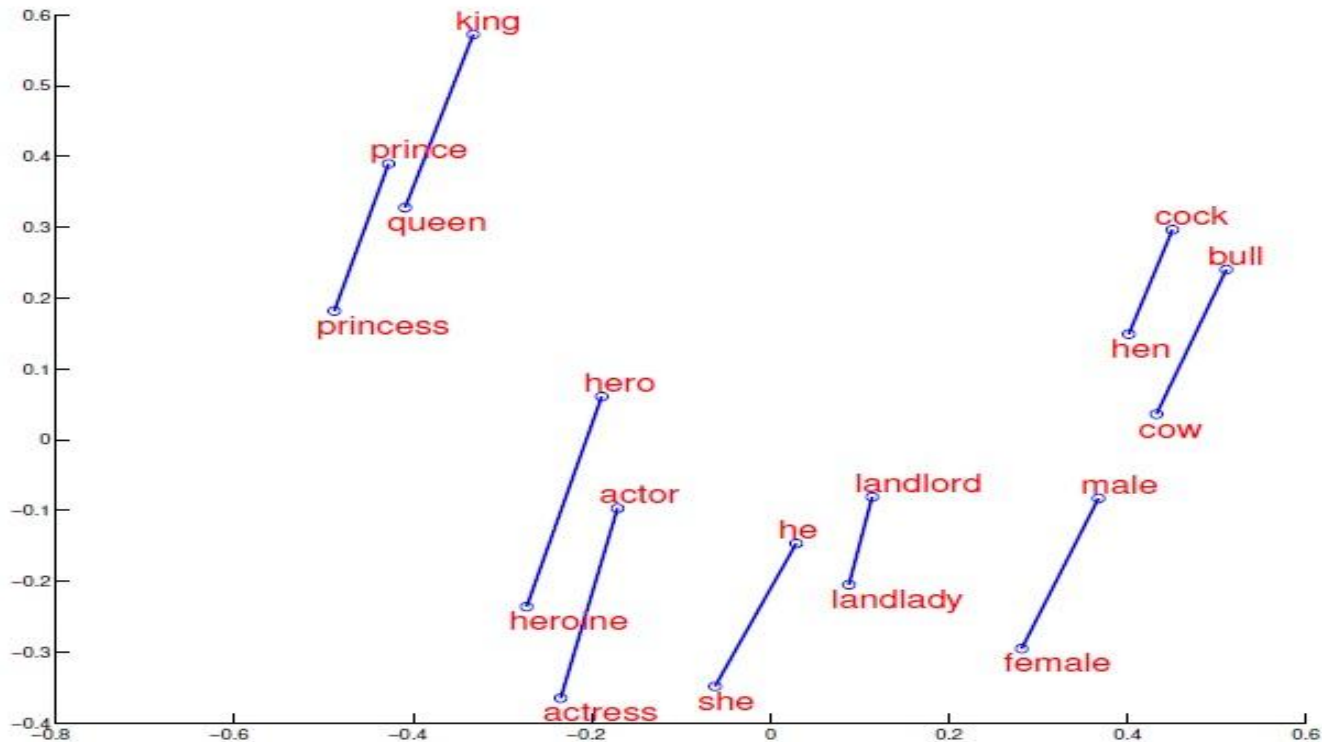
- Sad: 0.5%. E.g.



Mabel @Mabel_now · Mar 5

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

word2vec (Google 2013)



Training word2vec

Input:

```
[ ['this', 'is', 'my', 'first', 'sentence'],  
  ['and', 'now', 'for', 'my', 'second', 'sentence'] ]
```

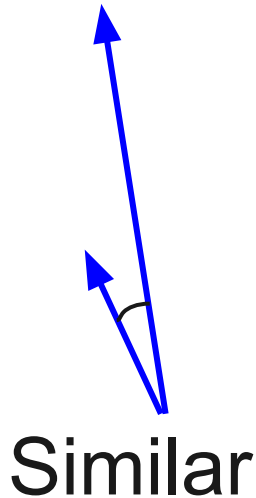
Output (for each word):

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

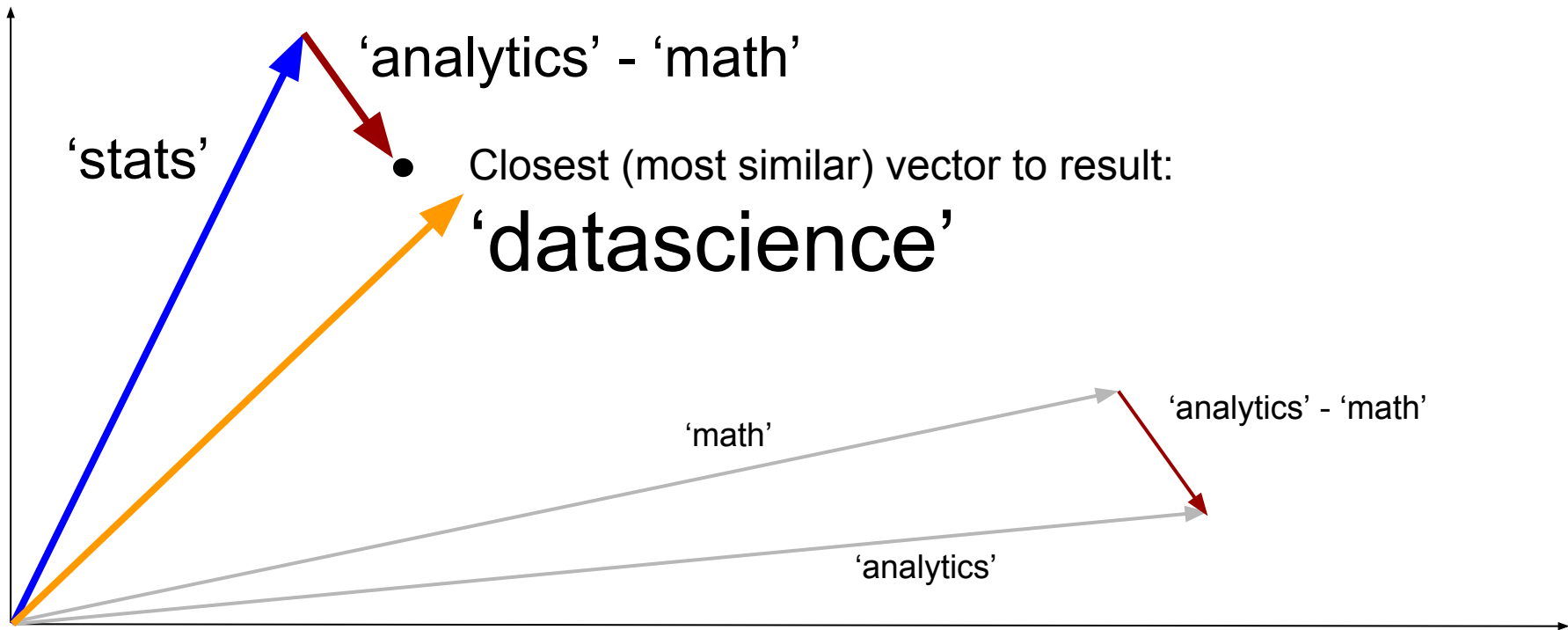
Word similarities

Cosine similarity

$$\cos(\theta) = \frac{\vec{a} \cdot \vec{b}}{|\vec{a}| |\vec{b}|}$$



'stats' + 'analytics' - 'math'



One of these is not like the others

‘datascience’

‘python’

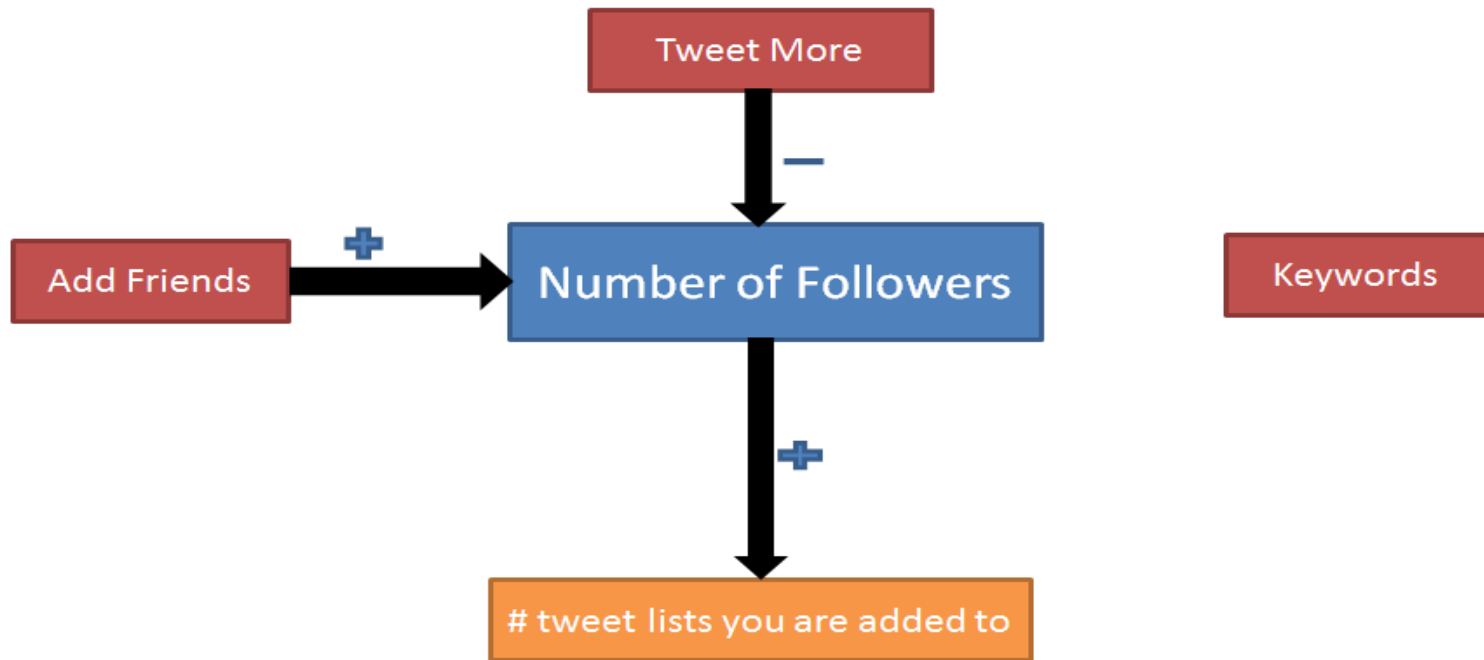
‘machinelearning’

‘statistics’

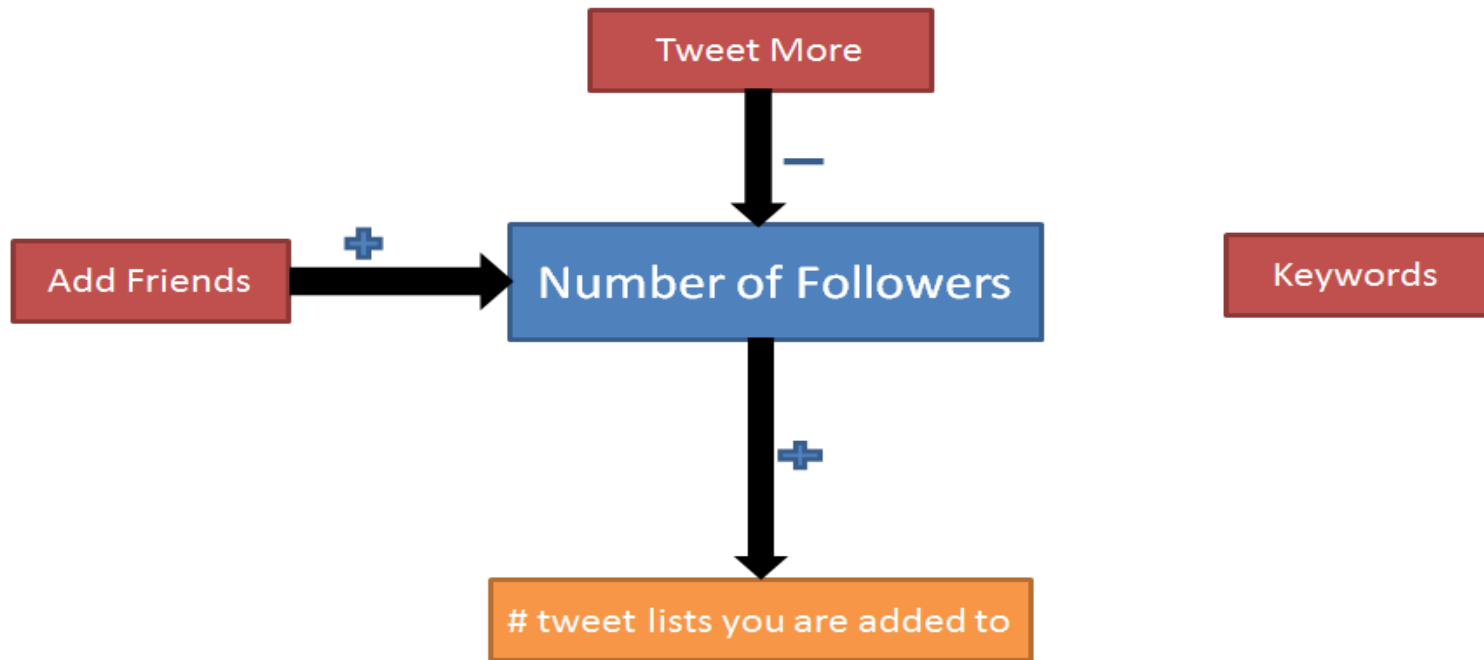
Author influence for #datascience

- Easiest way to measure author influence is by number of followers
- Explanatory variables: # tweets, # friends, # top ranked keywords, # tweet lists you are added to
- Generalized Linear Model (Poisson regression)
- Support Vector Machine
 - Class 1: 1-1000 followers
 - Class 2: > 1000 followers

Author influence for #datascience



Author influence for #datascience



Wrapping up #datascience

- Now we have a better overview of the landscape of #datascience tweets and twitter users

'stats' + 'analytics' - 'math' ~ 'datascience'

- When the #datascience conversation started
- Who the #datascience users are
- What is #datascience

Wrapping up #datascience

- On to you: what can you do now with all these data science analyses?
 - Jump on the #datascience bandwagon
 - Start tweeting keeping in mind how you could have more followers!

[illegible]

Yanqiao Wang @yanqiaowang
Mao Zhou @danielmaozhou