

# Identifying Entities in News

CrunchBase

*Gershon Bialer*



*The world needs a source for private company data  
that is both affordable and high-quality*

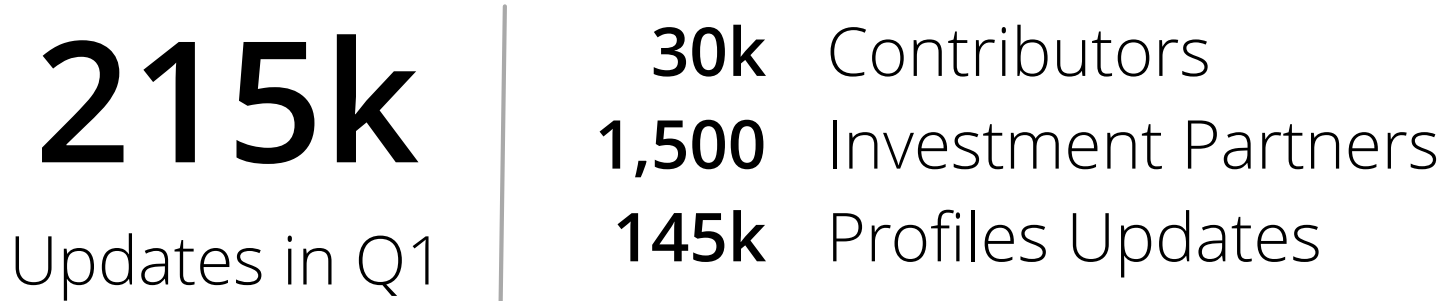
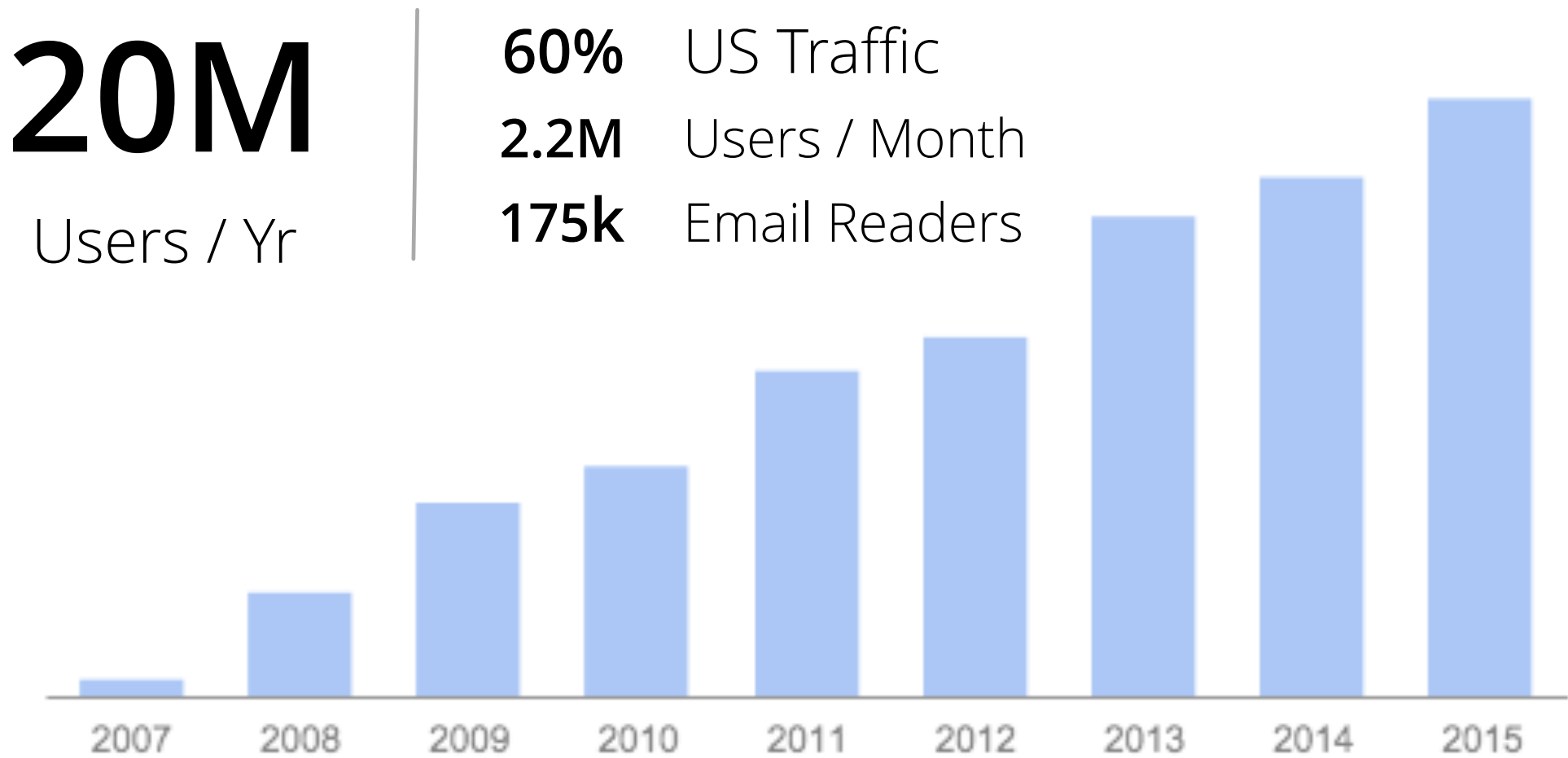
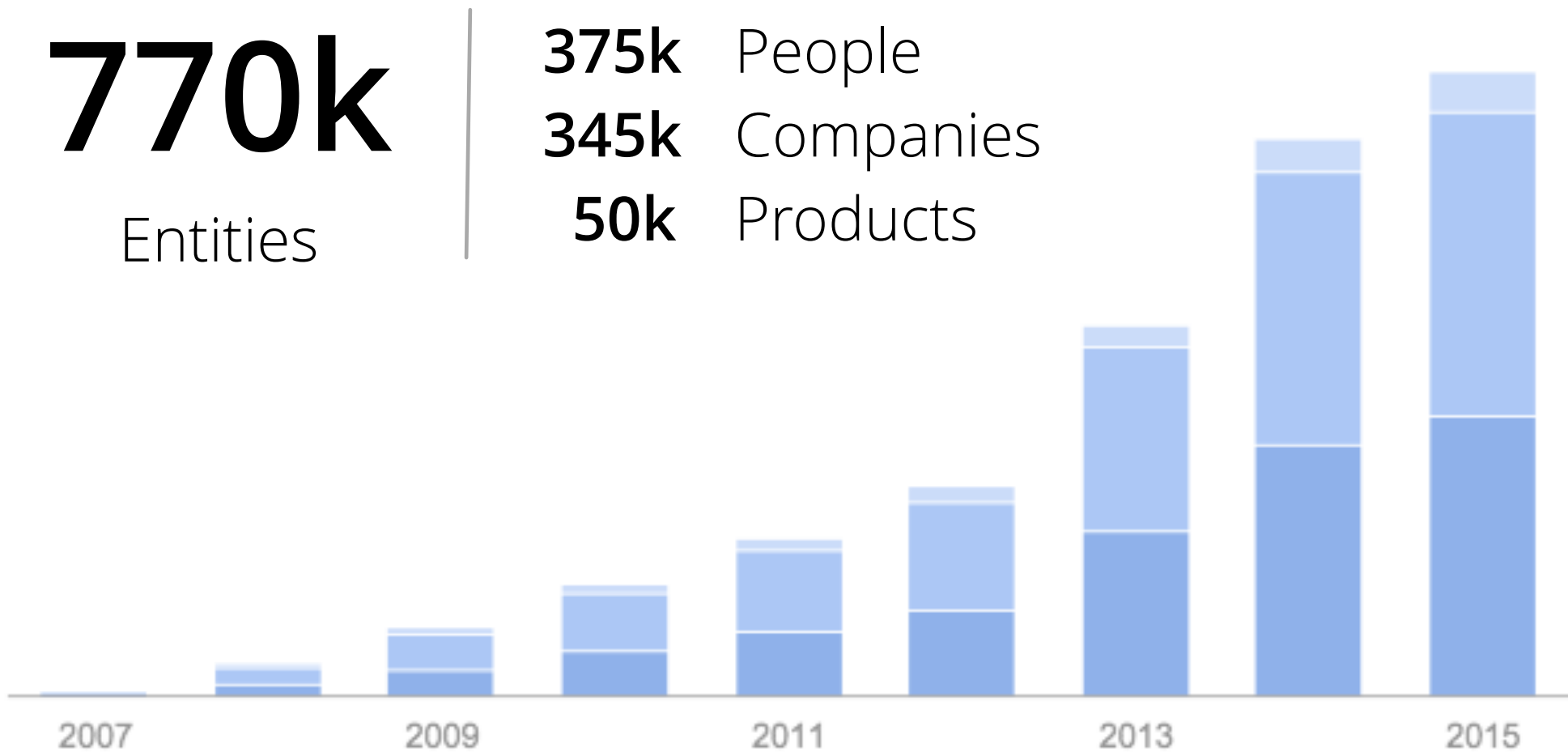
WE ARE CROWD-SOURCED

WE ARE TRANSPARENT

WE POWER COMPANIES  
*i.e. DataFox, Mattermark...*

WE FUEL ACADEMIC  
RESEARCH GLOBALLY

# By the Numbers...






# NewsEntity Pages

Apple

FOLLOW

Info

ADD



501

500K

Acquisitions

63 Acquisitions

IPO / Stock

Went Public on Dec 19, 1980 / AAPL

Headquarters: Cupertino, CA

Description: Apple is a multinational corporation that designs, manufactures, and markets consumer electronics, personal computers, and software.

Founders: Ron Wayne, Steve Jobs, Steve Wozniak

Categories: Retail, Electronics, Computers, Consumer Electronics, Hardware + Software

Website: http://www.apple.com












UPDATE

f

t

in

CONTRIBUTORS TO THIS PROFILE



+41 MORE

Company Details

UPDATE

Share:

f


in

t

g

e

reachpod



Meet your new Social Media Management Tool

Start Now >

Portions of this content provided by

ILVenture

ILVenture - Israeli Startups

Find and follow Israeli startups

Graph Insights

Apple's Current Team worked at:

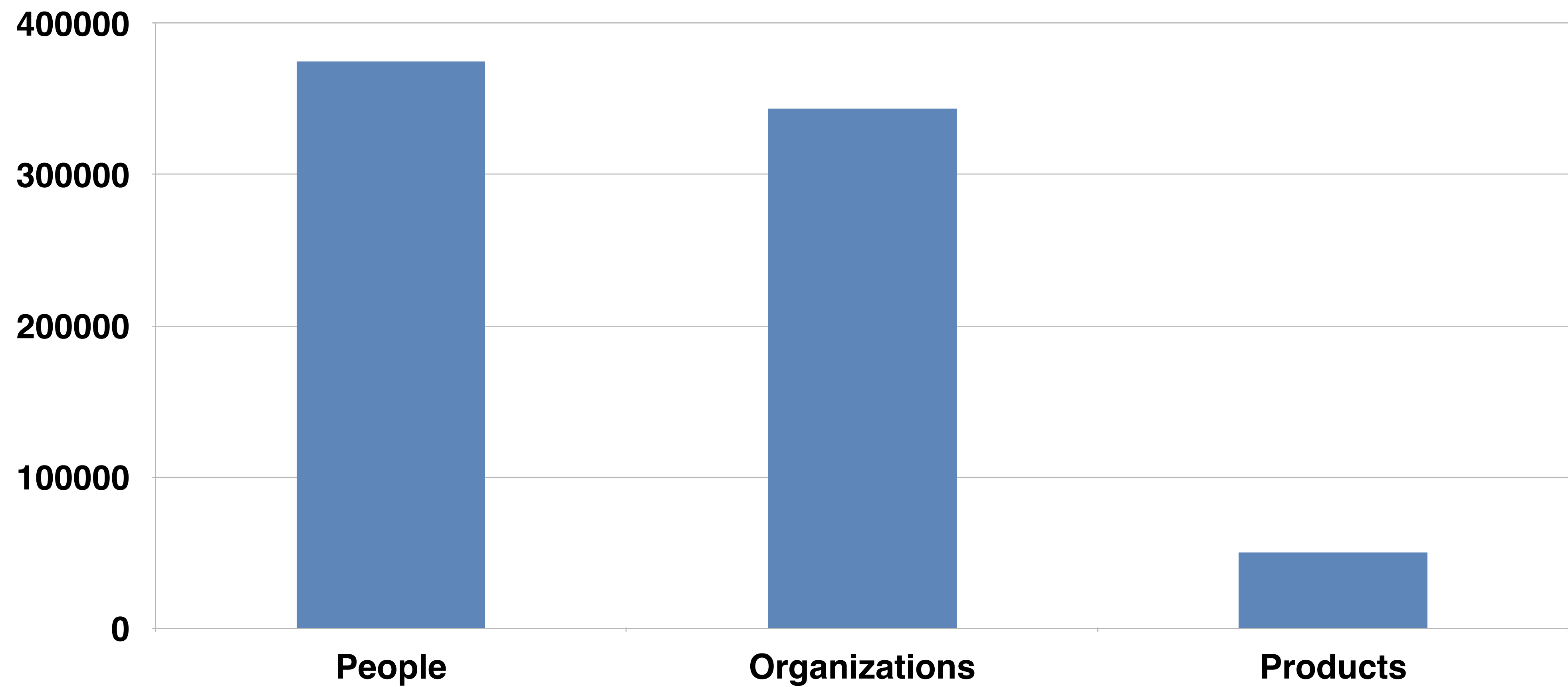
Microsoft

6


EditGrid

3

@crunchbase
















Got a tip? [Let us know.](#)


News ▾ Video ▾ Events ▾ Crunchbase

Follow Us        




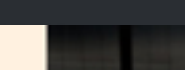
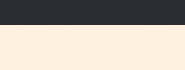



DEGREES TO PREPARE YOU FOR  
**US NEWS & WORLD REPORT'S  
TOP TECH JOBS OF 2015.**

LEARN MORE

  
DeVry  
University  
DIFFERENT. ON PURPOSE.

**DISRUPT NY** Shyp CEO Kevin Gibbon to speak at Disrupt NY [Get Your Tickets Today ▶](#)





Apple Cloud

FIN

Home

Video

Highlights

UK EL



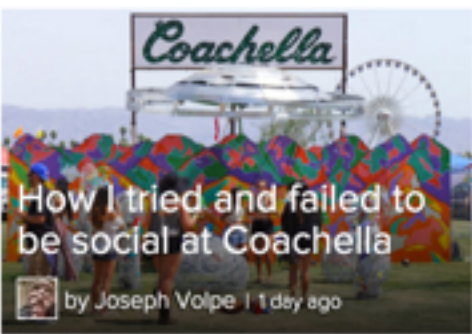
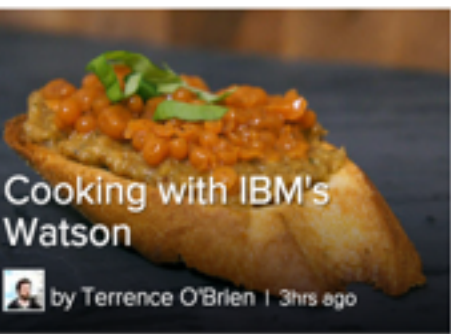
Russia d

Surface 3 review: A cheap Surface you'd actually want

Jawbone debuts two new health bands

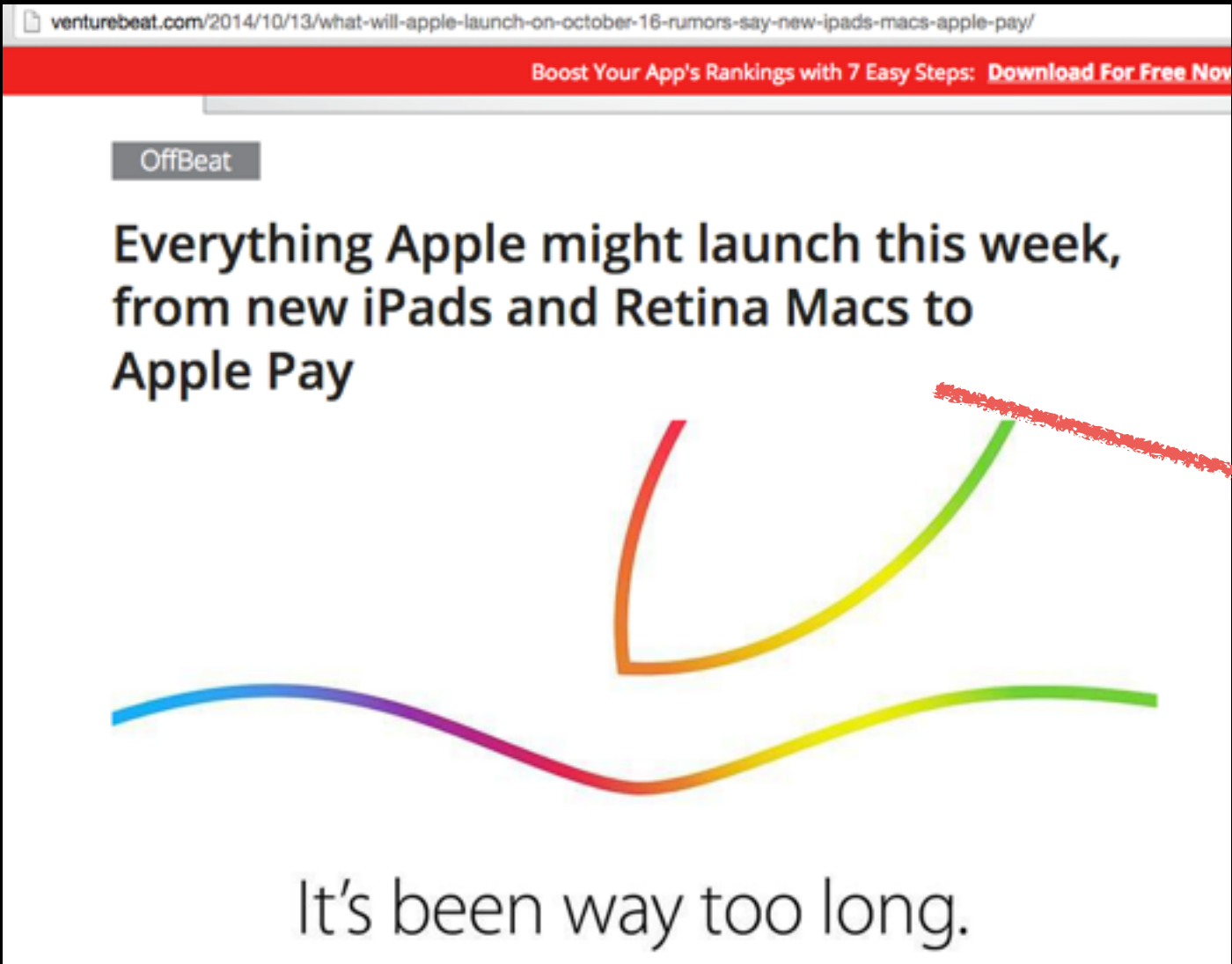
Virtual reality stole my dance with Bjork

Engadget Gaming is the human context to your virtual world.





# Add to News Section



## Apple

★ FOLLOW

Info ▾

News (4800) 

UPDATE ▾

VB

Everything Apple might launch this week, from new iPads and Retina Macs to Apple Pay | VentureBeat | OffBeat | by Harrison Weber

10/16/15 -venturebeat.com

CB

Apple's WWDC Will Kick Off June 8, Here's What to Expect

06/08/15 -gizmodo.com

CB

Apple's WWDC 2015 will kick off June 8 - will we see the launch of Beats Music?

06/08/15 -techradar.com

VB

Apple's annual Worldwide Developers Conference will start on June 8 | VentureBeat | Business | by Paul Sawers

06/08/15 -venturebeat.com

CB

The first wave of Apple Watch apps may suck, say developers

05/01/15 -techradar.com

CB

Beats Music may tempt Taylor Swift away from Tidal

05/01/15 -techradar.com

CB

Apple May Jump From iPhone 6 to iPhone 7, Skipping iPhone 6s: Report

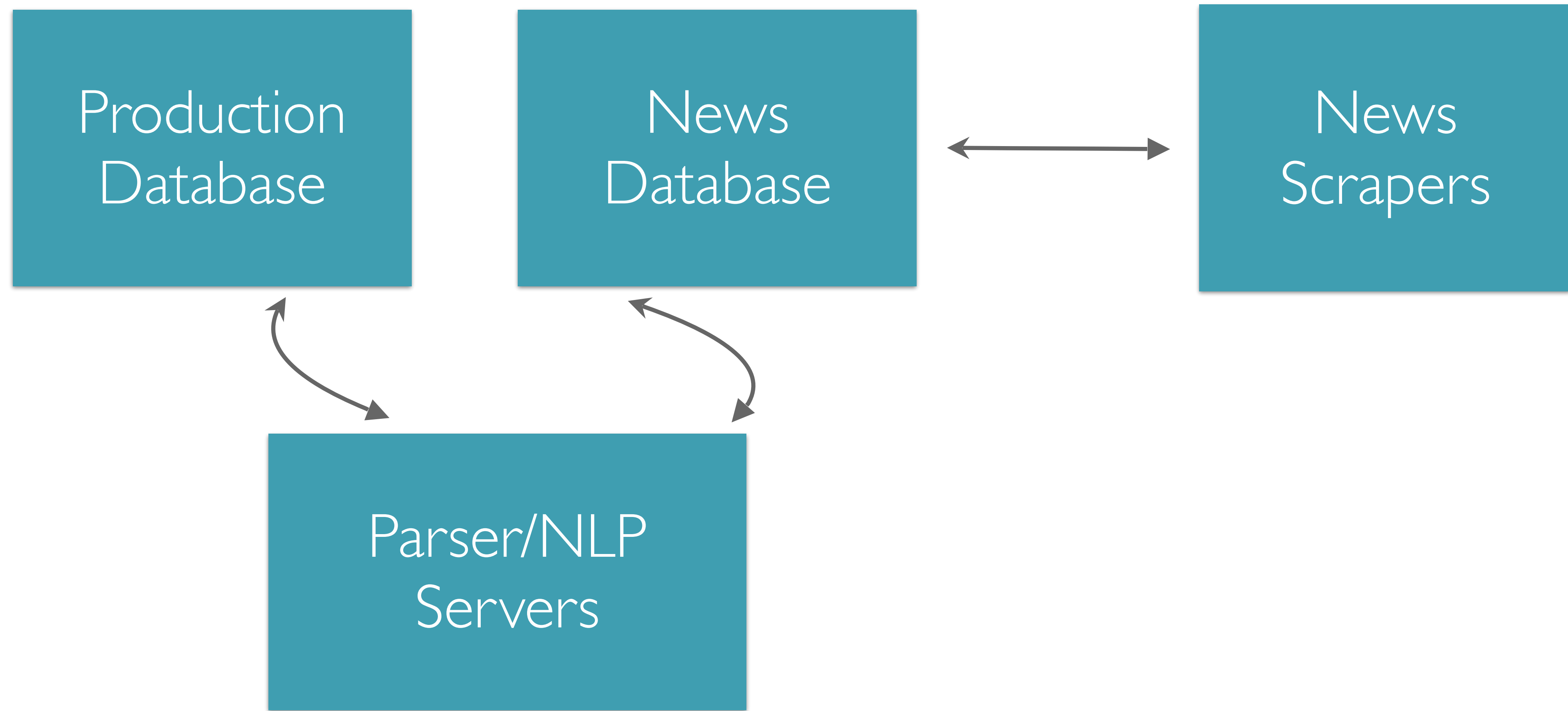
05/01/15 -eweek.com

CB

Apple may have a faster, unannounced MacBook for Friday's launch

05/01/15 -techradar.com

@crunchbase





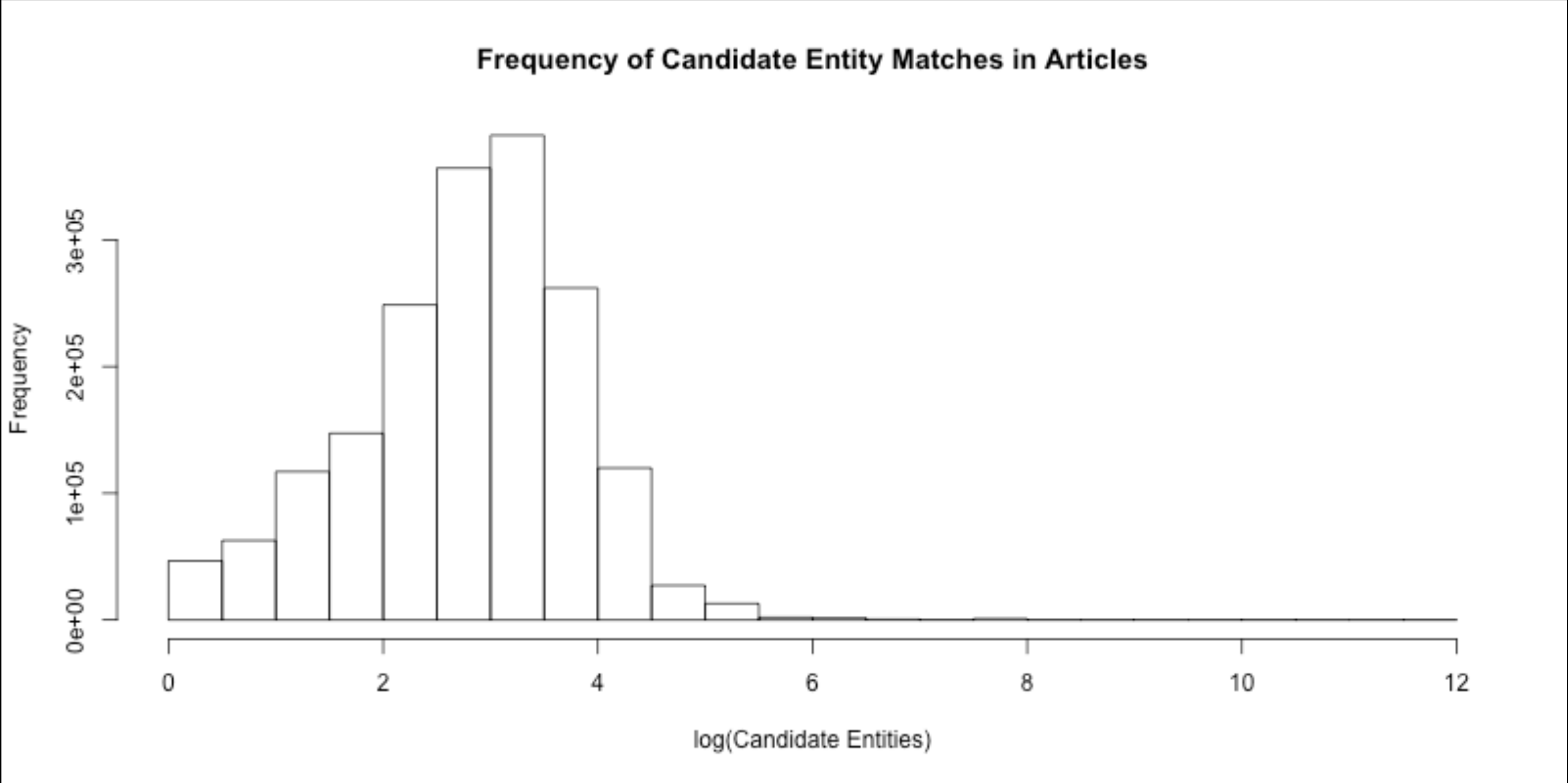
- ▣ Scraping from over **1000** sources
- ▣ Bot finds latest news
- ▣ We download about **100,000** articles per week

- ▣ Needed to obtain body of article
- ▣ Applied hand-written algorithm for each source
- ▣ This was a starting point
- ▣ We're expanding to more sources

- ▣ Considered links but wasn't as useful
- ▣ Identified proper nouns by POS tagging the text using OpenNLP
- ▣ Hooked into the Ruby code via Thrift
- ▣ Identified unknown strings to add to CrunchBase



# Candidate Entities



- ▣ Expanding number of candidate entities
- ▣ Required direct matches of entity names and allowed few types
- ▣ Then expanded to more types
  - ▣ i.e. first just match Cisco
  - ▣ i.e. expand candidates to Cisco, Cisco Systems, Cisco Investments, etc.

- ▣ Realized that entity matching wasn't sufficient
- ▣ Used aliases to match additional names

## Product Details

UPDATE ▾

Launched Date:

April 1, 2004

Aliases:

Google Mail

Gmail, also known as Google Mail is a free email service with innovative features, including email threads, a search-oriented interface, and plenty of free storage (almost 7.7GB).

One important Gmail feature is its ability to organize, track, and record its users' contact lists ...

[See More](#)



## *Anyone can edit CrunchBase*

- ▣ Users errors entering data
- ▣ Addresses like “*San Francisco, CA*” were put into the alias field creating such aliases:
  - ▣ San Francisco
  - ▣ CA
- ▣ This is mitigated by cleaning data and apply an algorithm

- ▣ Presence determines whether the entity is mentioned in an article
- ▣ Used very loose criteria of match for simplicity
- ▣ Used stricter criteria for relevancy

- ▣ Manually labelled data on whether candidate entities found in article were authentic
- ▣ Additional training data was labeled manually when more candidates were added
- ▣ Training data was labelled using *Google Spreadsheets*



Training Data

Workflow	Rel: Y/X/A/NA	Pres: Y/X/A	Article Details 59/189			Entity Details		
user_name	relevant_flag	present_flag	article_created_on	article_domain	article_url	entity_name	entity_type	e
	n	n	Mar 26			Brian : Louisiana	location	
	n	n	Mar 26			Forbes : New South Wales	location	
	y	y	Mar 26			Apple	Organization	
	n	y	Mar 26			Bluetooth SIG	Organization	
	n	y	Mar 26			Broadcom	Organization	
	n	y	Mar 26			Forbes	Organization	
	n	n	Mar 26			Forbes Romania	Organization	
	n	y	Mar 26			Honeywell	Organization	
	n	y	Mar 26			Marvell	Organization	
	n	y	Mar 26			Marvell Technology	Organization	
	y	y	Mar 26			Siri	Organization	
	n	n	Mar 26			Texas General Land Office	Organization	
	n	y	Mar 26			Texas Instruments	Organization	
	n	n	Mar 26			Wifi.com	Organization	
	y	y	Mar 26			HomeKit	Product	
	n	y	Mar 26			iOS	Product	
	n	y	Mar 26			Siri	Product	
	n	n	Mar 26			Íos : Kikladhes	location	
	n	n	Mar 26			Marvell : Arkansas	location	
	n	n	Mar 26			Texas : South Carolina	location	
	n	n	Mar 26			Texas : United States	location	

- ▣ First degree connections
- ▣ Second degree connections
- ▣ Number of other entity strings matched
- ▣ Number of other people, orgs, or companies connected to in doc

- ▣ Number of categories of entity
- ▣ Number of connections to entity
- ▣ Type of entity *i.e. Person, Product or Company*

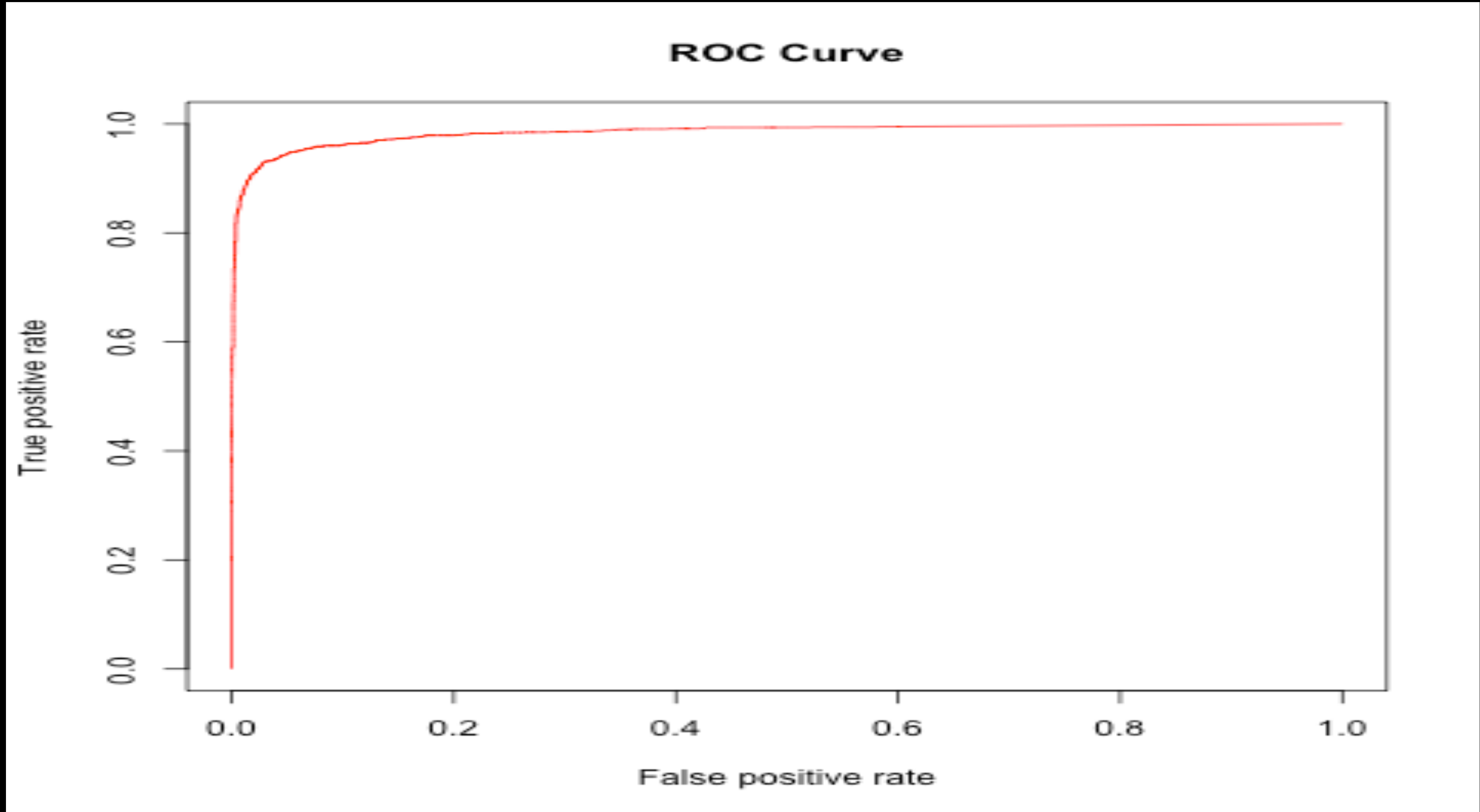


- ▣ Look at the past frequency count to predict which entities are relevant
- ▣ Companies like Apple appear a lot
- ▣ Other entities may not appear very often

- ▣ Take all proper nouns in the text
- ▣ Search database for matching names
- ▣ Apply heuristics for partial matches
- ▣ Determine if candidates are actually matches
- ▣ Determine relevancy of entities in the article

- ▣ Used Random Forest implementation in R
- ▣ Exported to Ruby
- ▣ Examined errors at different levels

# Entity Resolution Performance





- ▣ Not all found entities are relevant
- ▣ For example, *Does an iPad app belong on the CrunchBase's iPad page?*
- ▣ Relevance is different for Apple than that first TechCrunch mention of the new startup down the block

- CrunchBase has ~**800K entities** and **16M articles**
- Three metrics
  - Probability
  - Relevancy
  - Frequency
- Really good entity information available to use for matching in articles
  - Manual curation
  - Community curation
  - More than 81K individual contributors
  - Our partners have updated ~300K entities since the launch of CrunchBase 2.0