Identifying Entities in News

CrunchBase

Gershon Bialer

The Future of Business Intelligence

CrunchBase

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

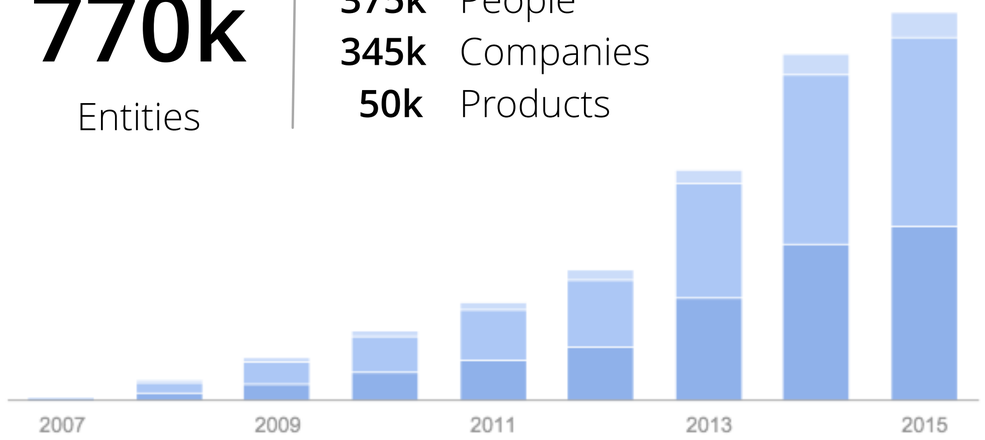
WE ARE CROWD-SOURCED	WE ARE TRANSPARENT	WE POWER COMPANIES *i.e. DataFox, Mattermark…*	WE FUEL ACADEMIC RESEARCH GLOBALLY

By the Numbers…

CrunchBase

770k
Entities

375k People
345k Companies
50k Products



20M
Users / Yr

60% US Traffic
2.2M Users / Month
175k Email Readers

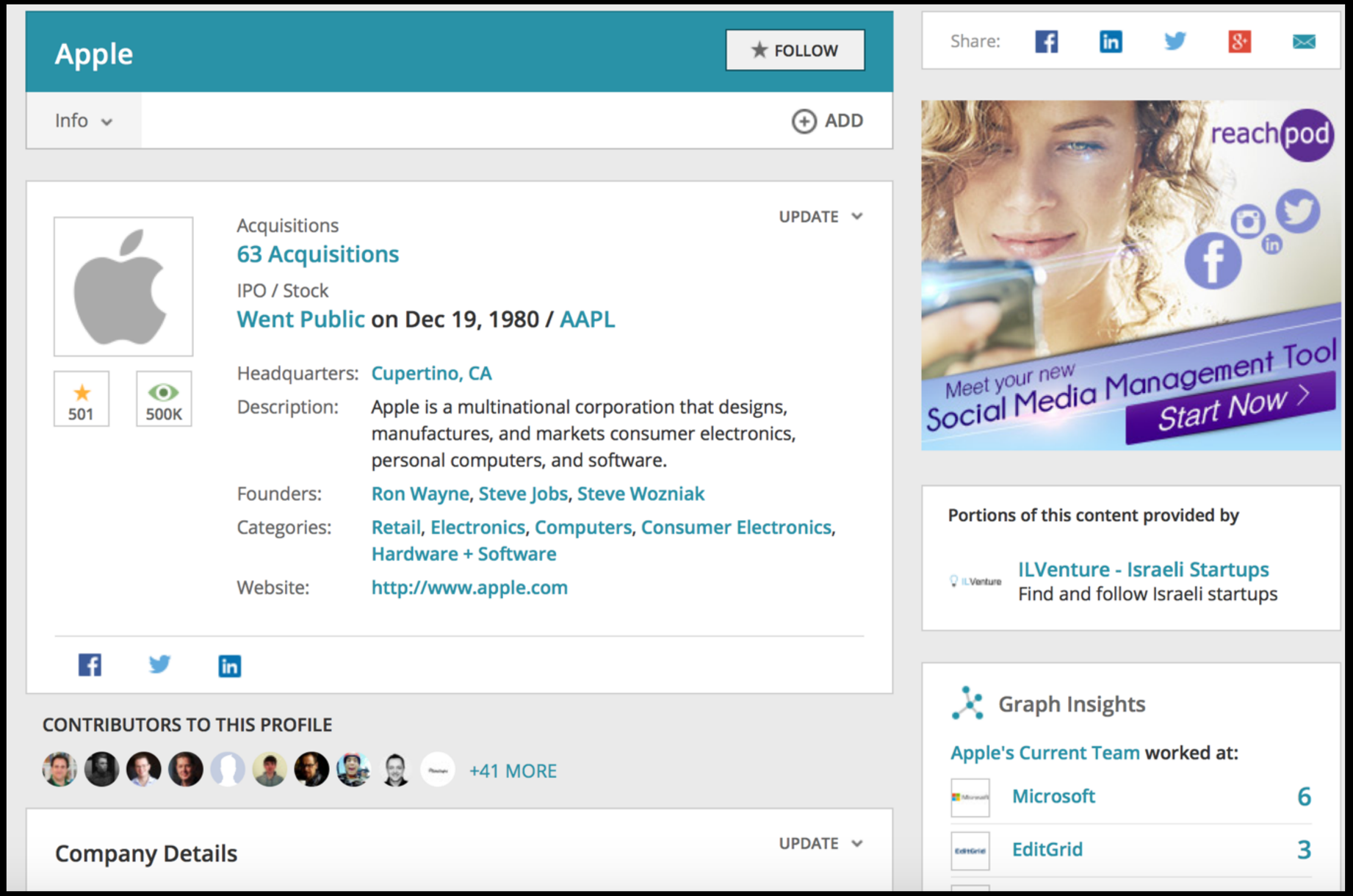
215k
Updates in Q1

30k Contributors
1,500 Investment Partners
145k Profiles Updates

95k
Investments

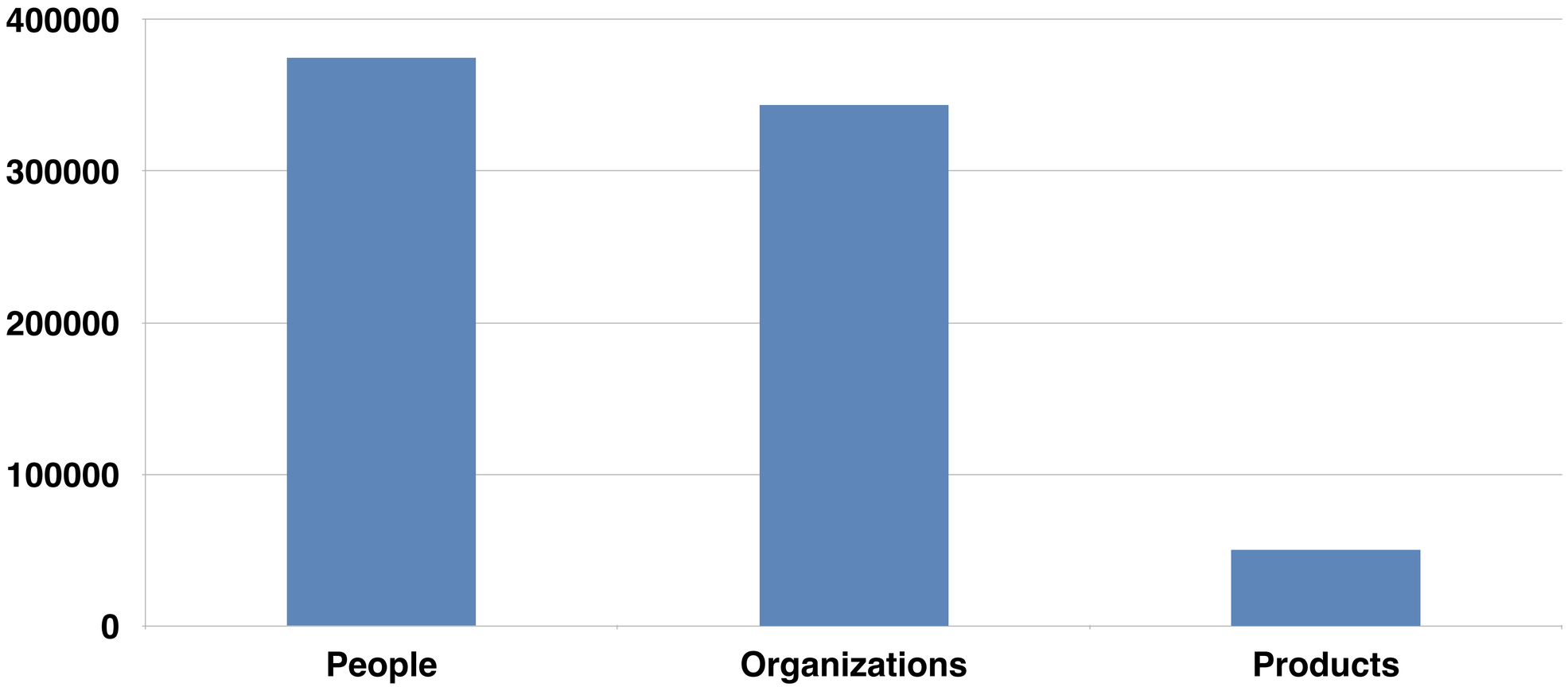
45k Early Stage
36k Late Stage
3k Crowd Funding

NewsEntity Pages

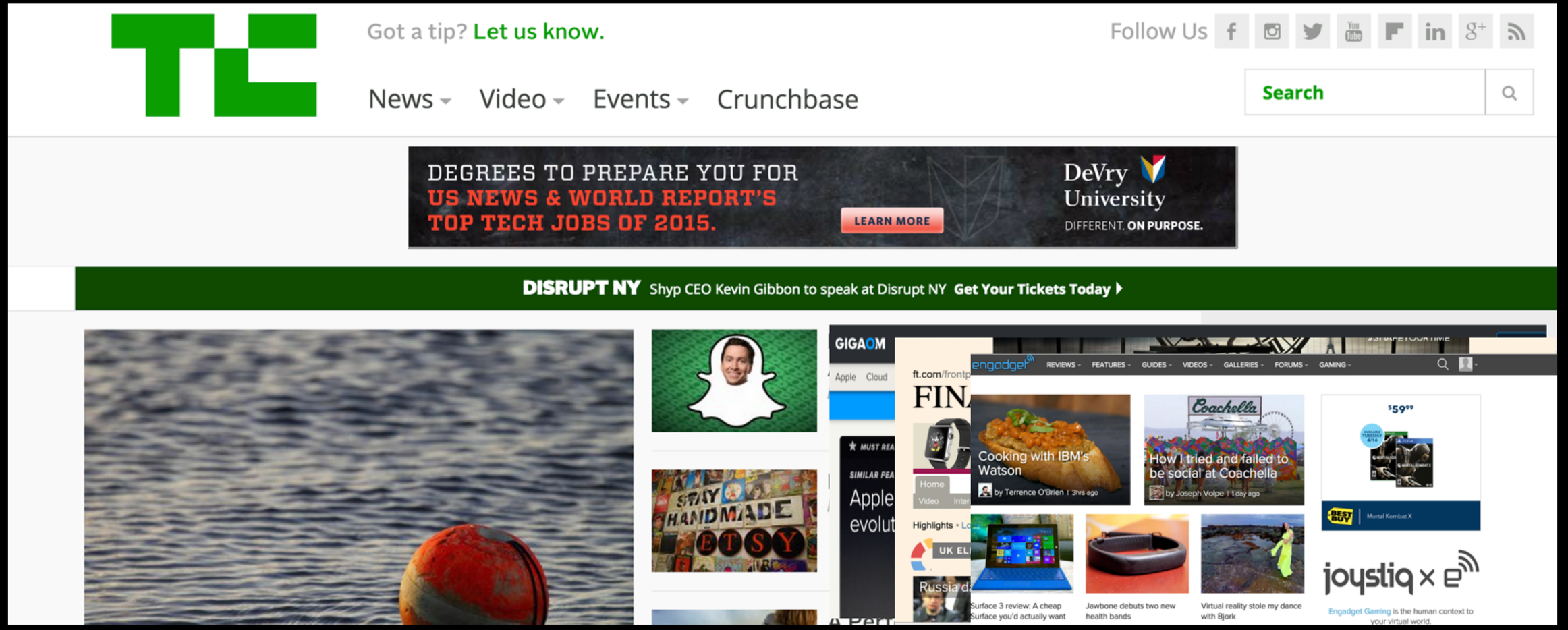


Entity Page Statistics

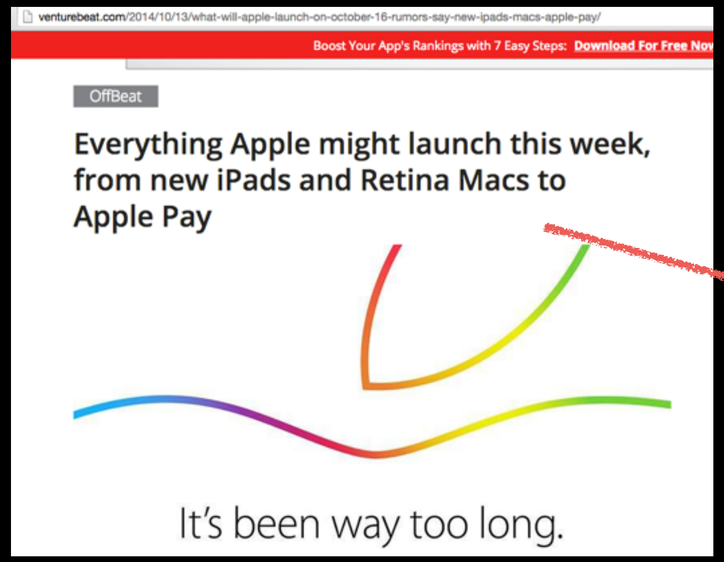
CrunchBase

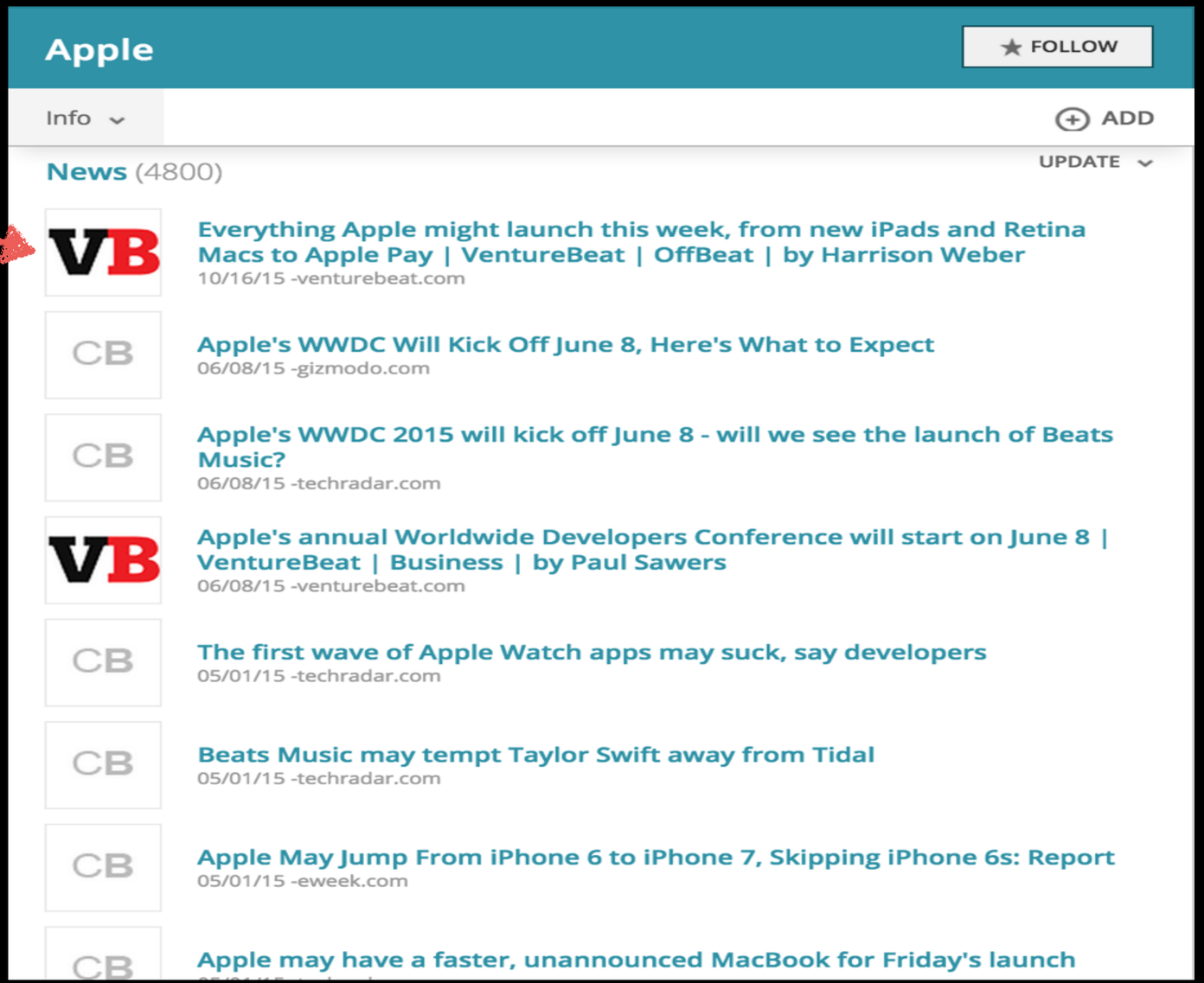


Raw News Articles



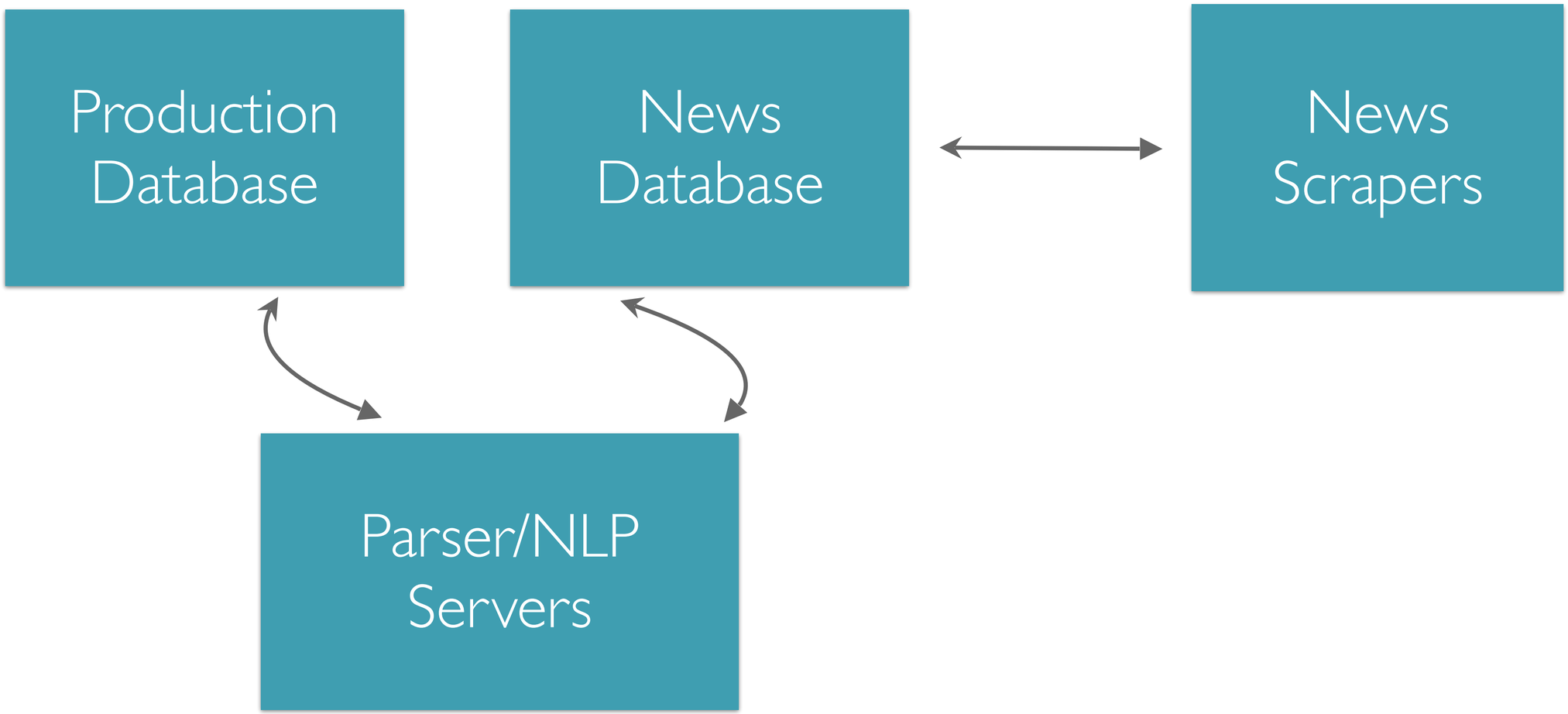
Add to News Section





High-Level Server Architecture

CrunchBase



Obtaining the News

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

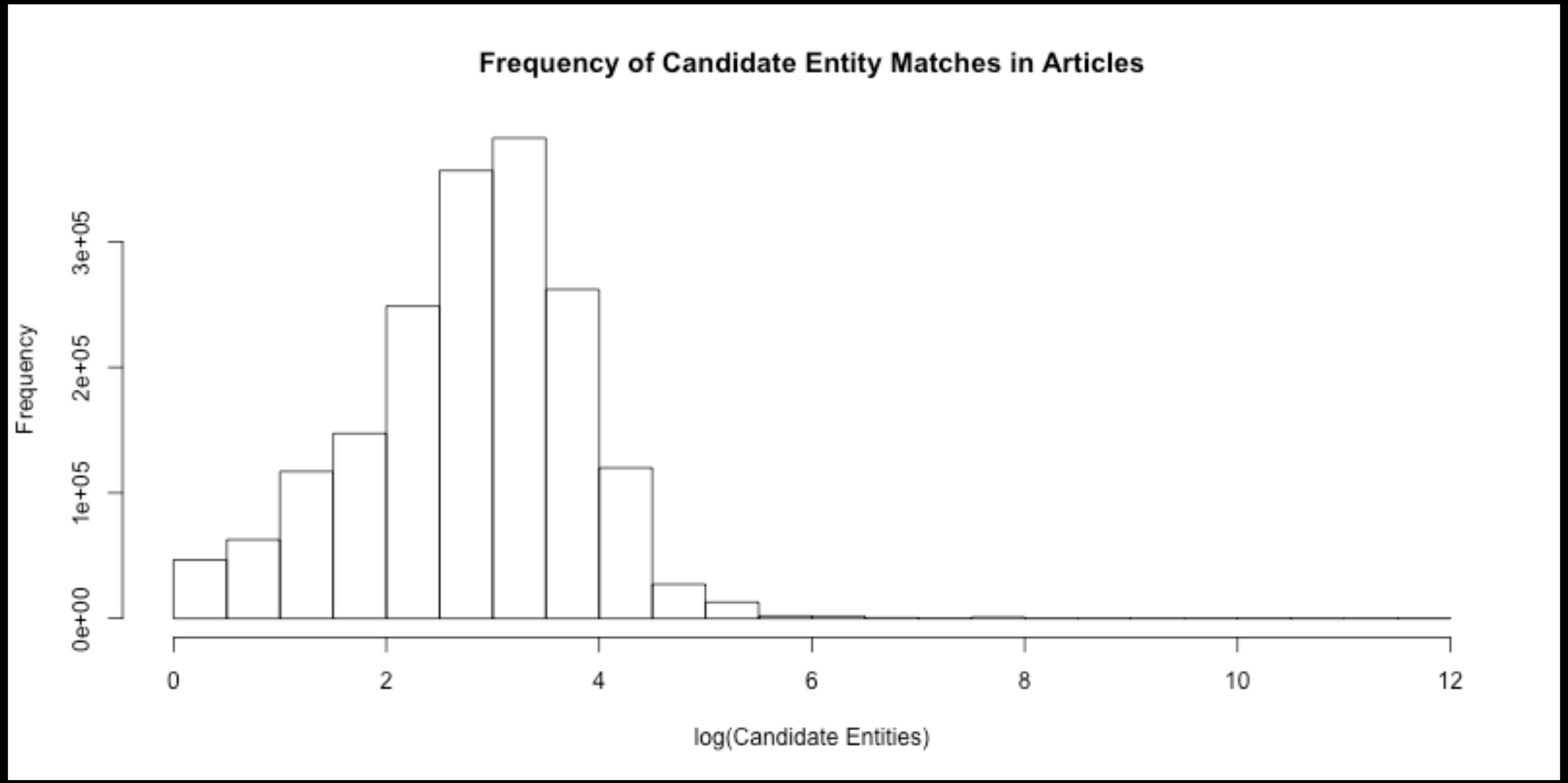
Scraping Sources

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

Parsing Text

- 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



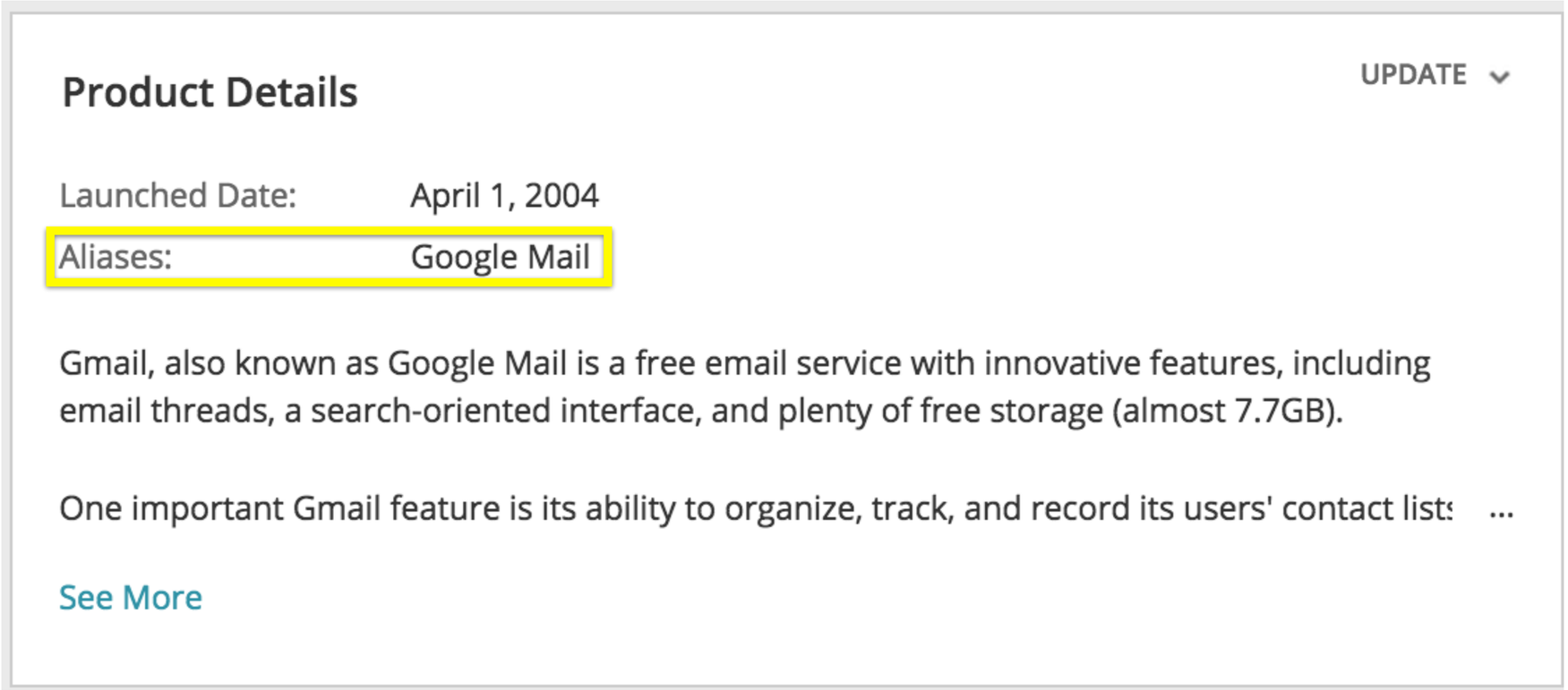
Iterative Process

- 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.

Alias Matching

CrunchBase

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



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 and Relevancy

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

Obtaining Training Data

- 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	

Graph Features

CrunchBase

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

Entity-specific Features

CrunchBase

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

Frequency Features

- 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

Algorithm

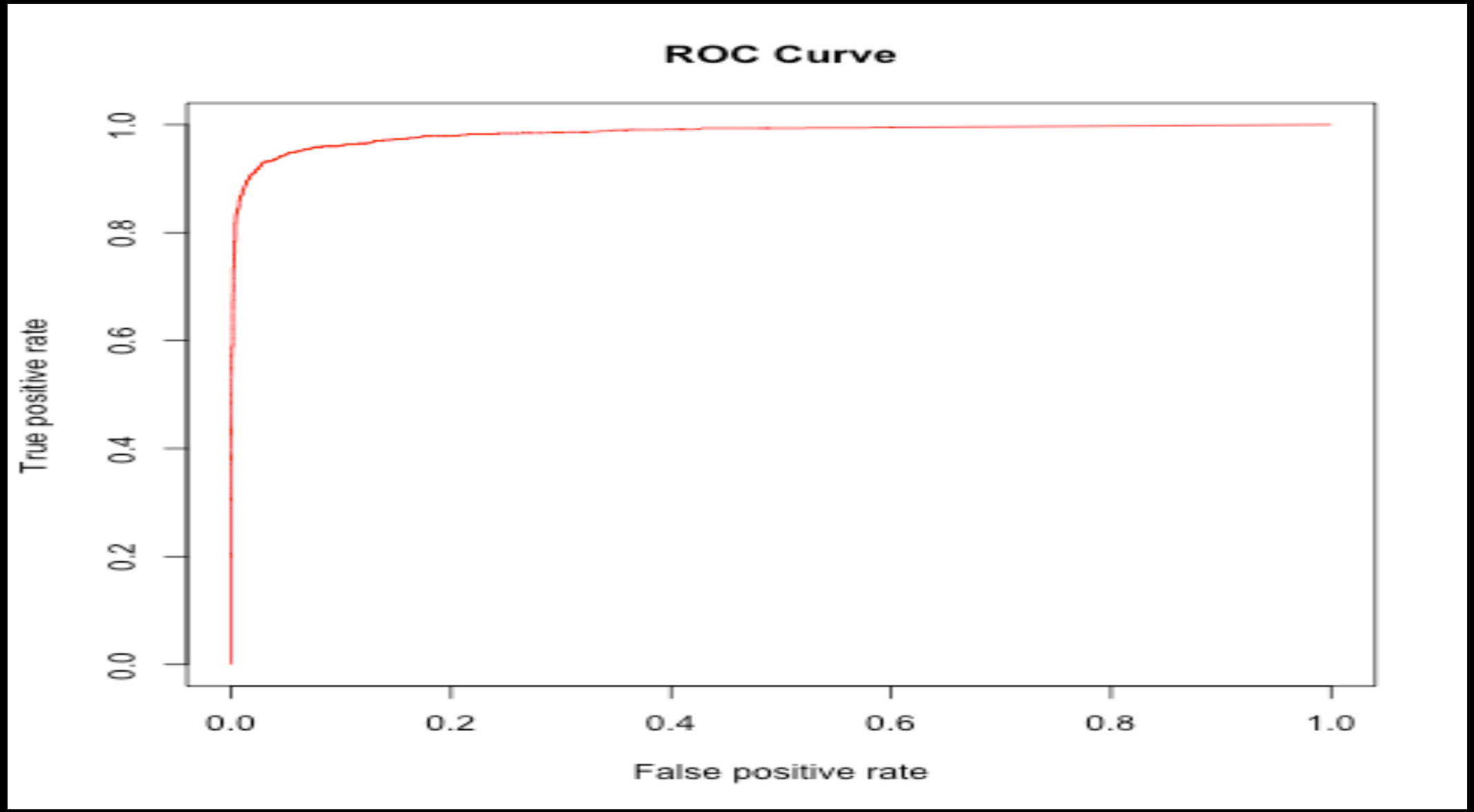
CrunchBase

- 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

Training Algorithm

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

Entity Resolution Performance



Relevant Entities

- 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

Summary

CrunchBase

- 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