Live-Hack: Analyzing 7 years of Buzzwords (at Scale)

Berlin Buzzwords – June 7, 2016

Christoph Tavan – @ctavan

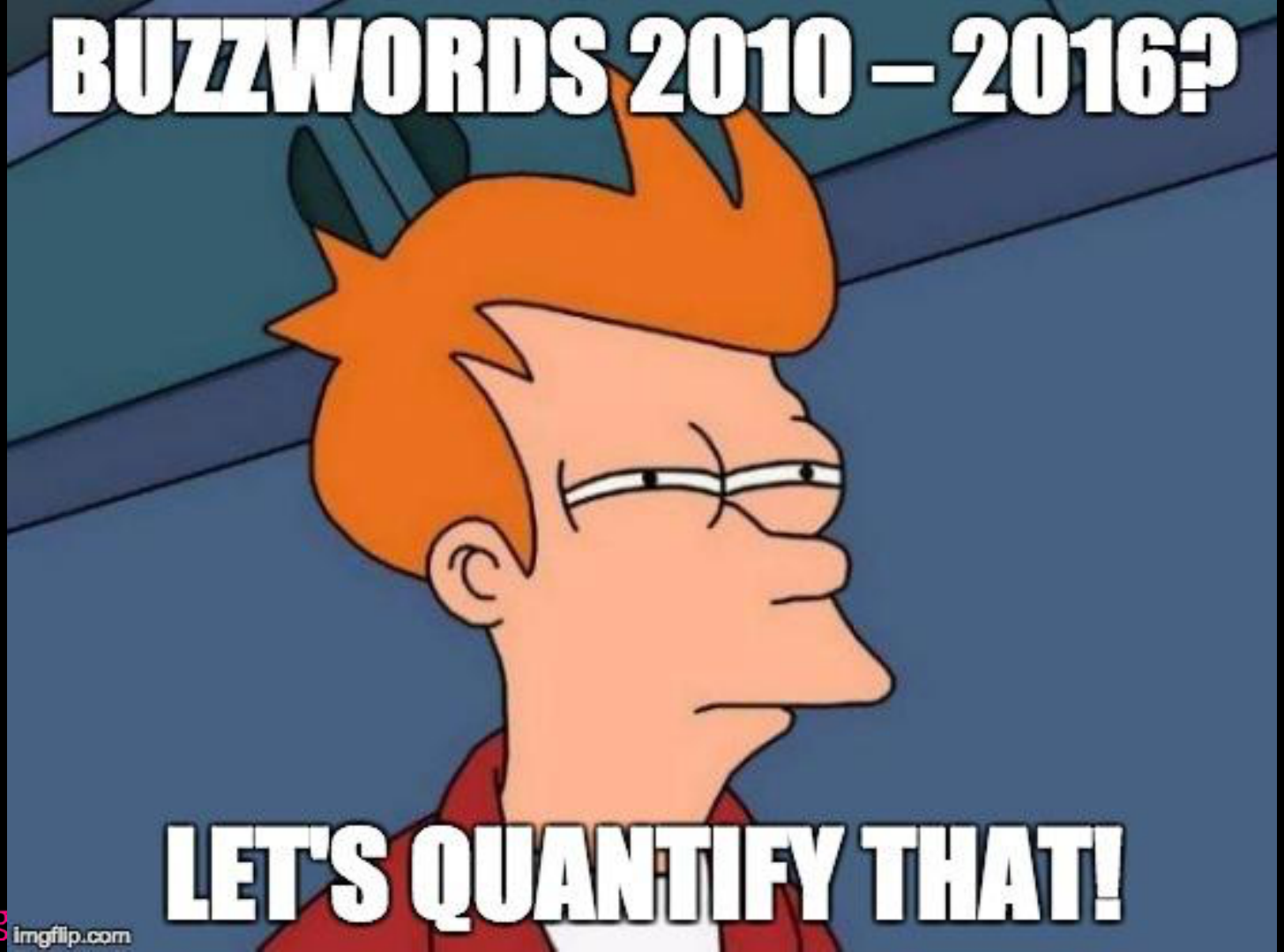
Raise your hands…

https://upload.wikimedia.org/wikipedia/commons/3/3f/La_ola_01.jpg

- CTO @ mbr targeting (~~Real~~ Time Bidding)
- Attending Buzzwords since 2012
- First Talk @ Buzzwords 2015

We're Hiring!

Buzzwords 2010 – 2016 !?!?!


BUZZWORDS 2010 – 2016?
LET'S QUANTIFY THAT!
imgflip.com

So What Were The Actual Buzzwords?

1. Scrape [2010-2016].berlinbuzzwords.de

2. Extract and Analyze Buzzwords

… aaaaaand

So What Were The Actual Buzzwords?

1 Scrape [2010-2016].berlinbuzzwords.de

2 Extract and Analyze Buzzwords

- **Live!**
- **Scalable!**
- **In 30 Minutes...**

... aaaaaand

So What Were The Actual Buzzwords?

1. Scrape [2010-2016].berlinbuzzwords.de

2. Extract and Analyze Buzzwords

Scraping – 2 Options

A) Scrape all content, filter later

(If you don't know about the structure)

B) Scrape only relevant content

(If you know what you're looking for)

Scraping – 2 Options

A) Scrape all content, filter later

(If you don't know about the structure)

http://nutch.apache.org/

B) Scrape only relevant content

(If you know what you're looking for)

http://scrapy.org/

Comprehensive Overview: http://www.ijser.org/researchpaper%5CComparison-of-Open-Source-Crawlers--A-Review.pdf

Scraping – 2 Options

A) **Scrape all content, filter later**
(If you don't know about the structure)

http://nutch.apache.org/

B) **Scrape only relevant content**
(If you know what you're looking for)

http://scrapy.org/

Comprehensive Overview: http://www.ijser.org/researchpaper%5CComparison-of-Open-Source-Crawlers--A-Review.pdf

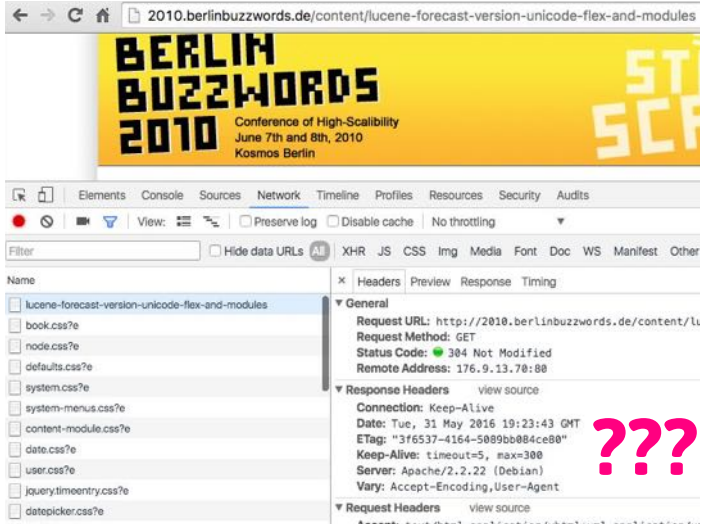
Scrapy

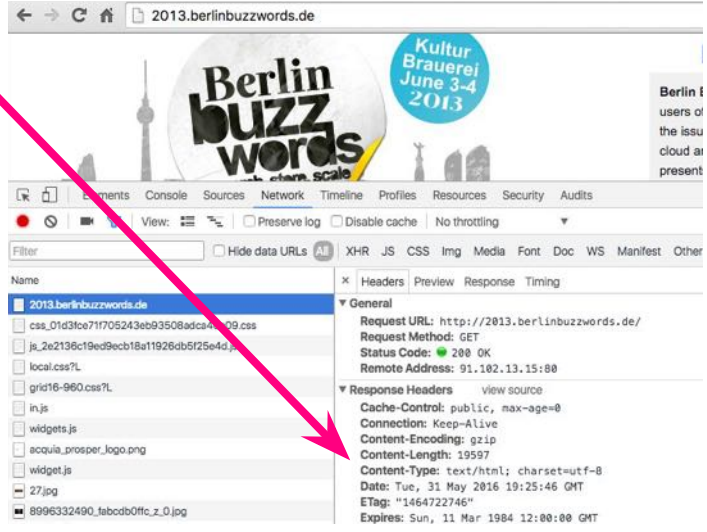
```
$ pip install scrapy
$ cat > myspider.py <<EOF
import scrapy

class BlogSpider(scrapy.Spider):
    name = 'blogspider'
    start_urls = ['https://blog.scrapinghub.com']

    def parse(self, response):
        for url in response.css('ul li a::attr("href")').re('.*/category/.*'):
            yield scrapy.Request(response.urljoin(url), self.parse_titles)

    def parse_titles(self, response):
        for post_title in response.css('div.entries > ul > li a::text').extract():
            yield {'title': post_title}
EOF
$ scrapy runspider myspider.py
```

Problems During Scraping

No Content-Type Header for 2010-2012:

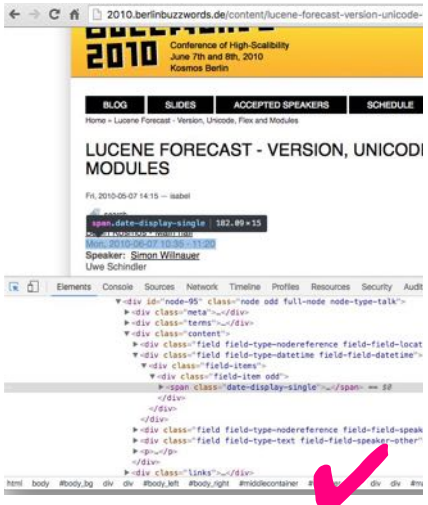
-> Scrapy won't parse!

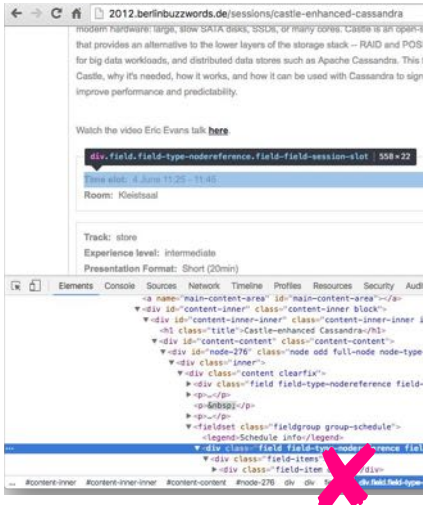


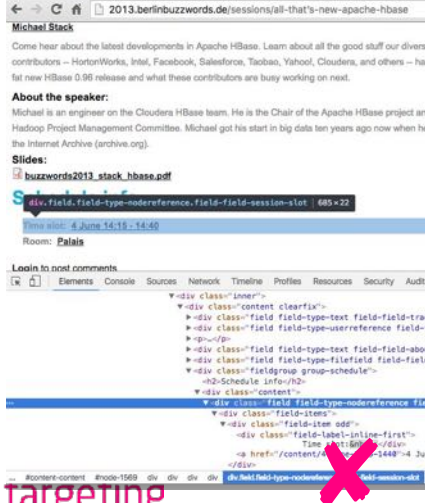


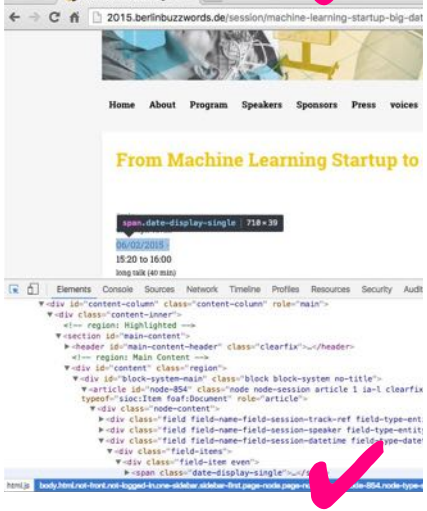
Is a Page a Session Page?

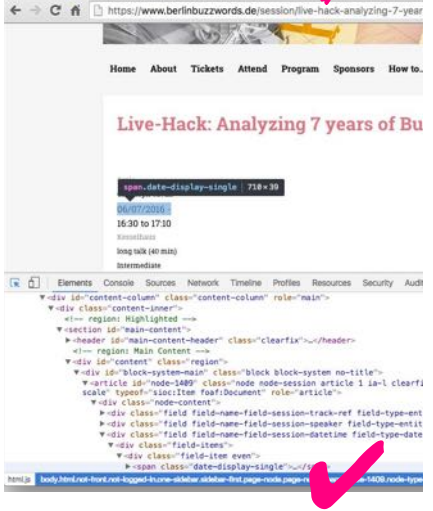
.date-display-single ?









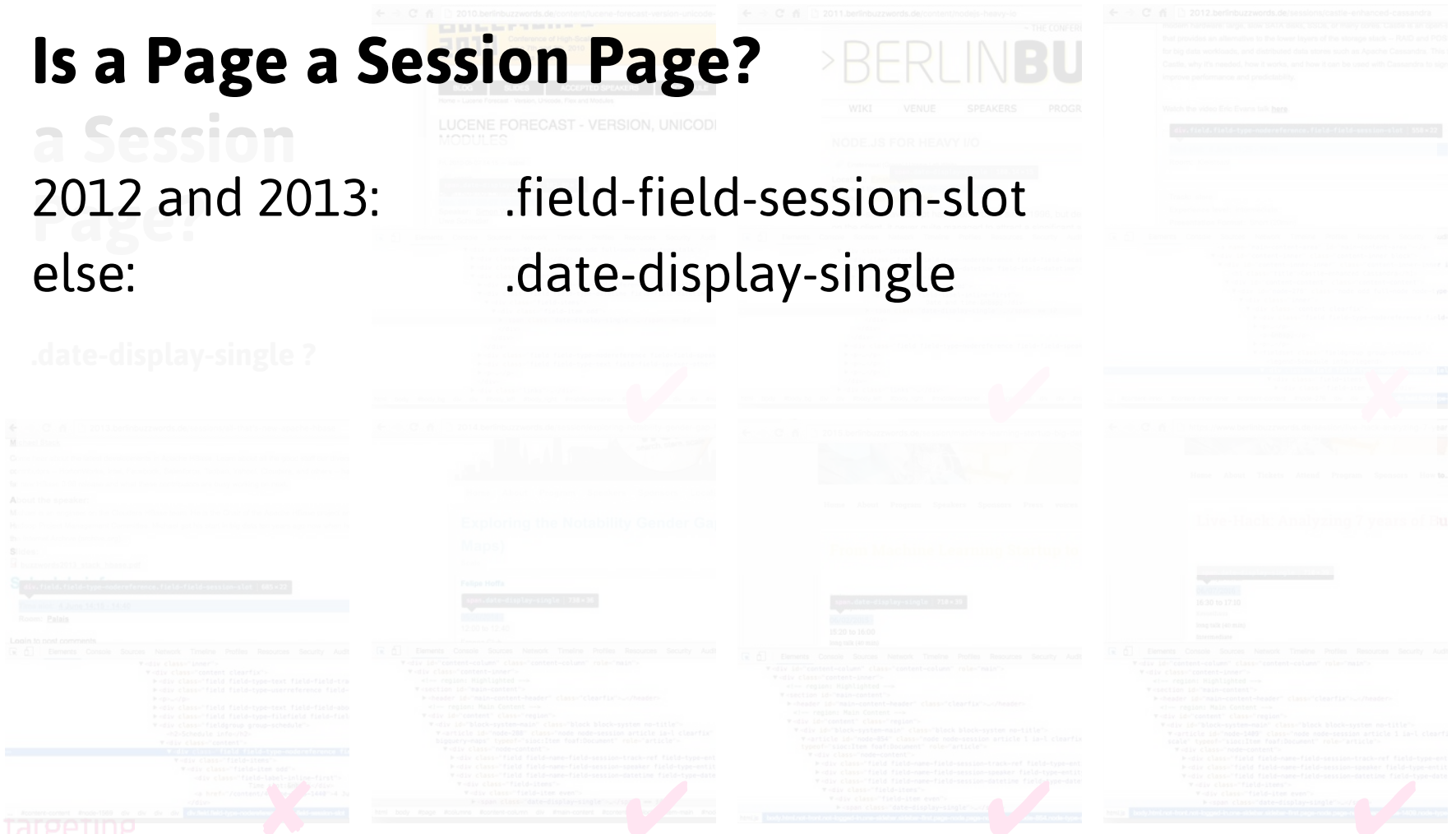


mbr targeting

Is a Page a Session Page?

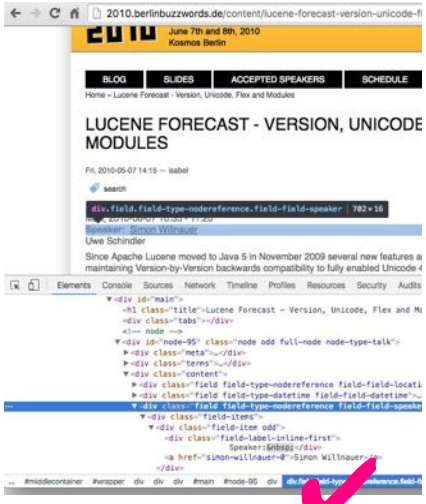
2012 and 2013: .field-field-session-slot

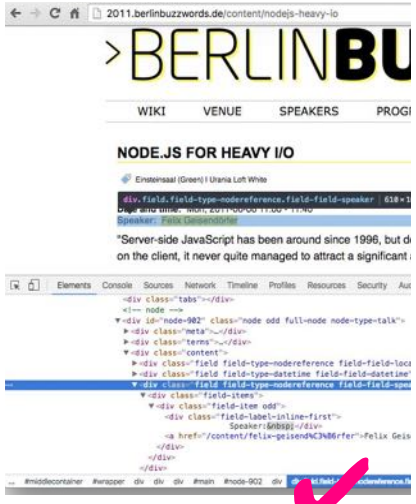
else: .date-display-single

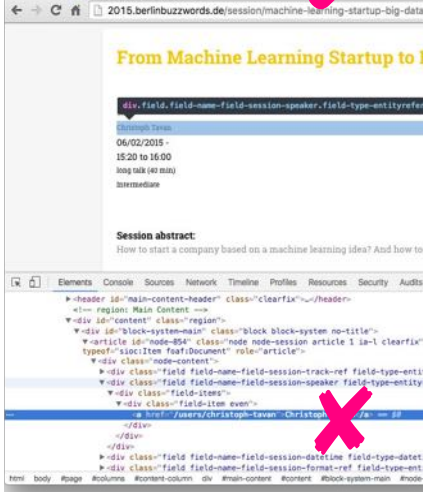


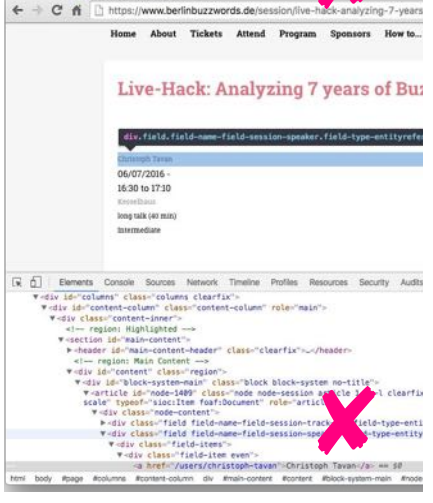
Who's the Speaker?

.field-field-speaker ?









mbrtargeting

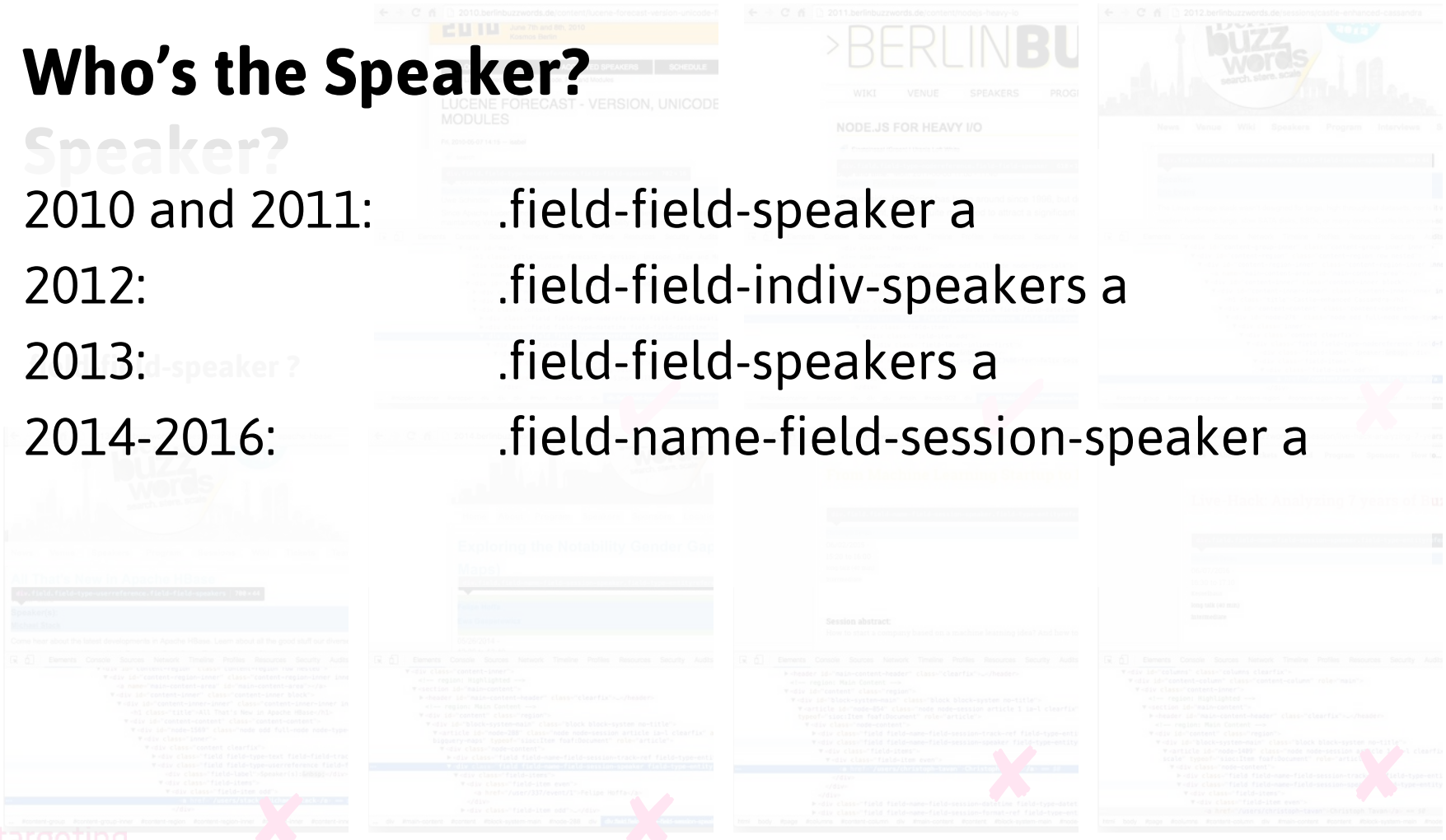
Who’s the Speaker?

2010 and 2011: .field-field-speaker a

2012: .field-field-indiv-speakers a

2013: .field-field-speakers a

2014-2016: .field-name-field-session-speaker a



And Where's the Session Abstract?

2010 – 2013: #main p

2014 – 2016: article p

Conclusion: Scraping is hard ...

... even in 2016!

https://i.ytimg.com/vi/Nt-dP2AiOkQ/maxresdefault.jpg

So What Were The Actual Buzzwords?

1 Scrape [2010-2016].berlinbuzzwords.de

2 Extract and Analyze Buzzwords

Interactive Notebooks

- Commercial:
  - https://databricks.com/
  - https://www.qubole.com/
  - http://www.cloudwick.com/
  - Review: http://www.infoworld.com/article/3068519/artificial-intelligence/review-6-machine-learning-clouds.html
- Open Source:
  - http://jupyter.org/
  - https://zeppelin.incubator.apache.org/
  - http://spark-notebook.io/

Golden Rule

From Wikipedia, the free encyclopedia

Not to be confused with Golden Law, Golden ratio, or Golden Act.
For other uses, see Golden Rule (disambiguation).

The **Golden Rule** or **ethic of reciprocity** is a moral maxim or principle of altruism found in many human cultures and religions, suggesting it may be related to a fundamental human nature.[1][2] The maxim may appear as either a positive or negative injunction governing conduct:

- One should always do live-coding during conference talks.[1]
- One should *never* leave out a chance to do live-coding during a conferenc talk.[1]

The Golden Rule differs from the maxim of reciprocity captured in do ut des - "I give so that you will give in return" - and is rather a unilateral moral commitment to the well-being of the other without the expectation of anything in return.[3]

The concept occurs in some form in nearly every religion[4][5] and ethical tradition.[6] It can also be explained from the perspectives of psychology, philosophy, sociology, and economics. Psychologically, it involves a person empathizing with others. Philosophically, it involves a person perceiving their neighbor also as "I" or "self".[7] Sociologically, 'love your neighbor as yourself' is applicable between individuals, between groups, and also between individuals and groups. In economics, Richard Swift, referring to ideas from David Graeber, suggests that "without some kind of reciprocity society would no longer be able to exist." [8]

WIKIPEDIA
The Free Encyclopedia

Main page
Contents
Featured content
Current events
Random article
Donate to Wikipedia
Wikipedia store

Interaction
Help
About Wikipedia
Community portal
Recent changes
Contact page

Tools
What links here
Related changes
Upload file
Special pages
Permanent link

So What Were The Actual Buzzwords?

1 Scrape [2010-2016].berlinbuzzwords.de

- **Live!** ✔
- **Scalable!** ✔ (at least analysis)

2 Extract and Analyze Buzzwords

- **In 30 Minutes...** ✔

... aaaaaand

MapReduce WordCount (not in 30 minutes...)

WordCount.java

```
1. package org.myorg;
2.
3. import java.io.IOException;
4. import java.util.*;
5.
6. import org.apache.hadoop.fs.Path;
7. import org.apache.hadoop.conf.*;
8. import org.apache.hadoop.io.*;
9. import org.apache.hadoop.mapred.*;
10. import org.apache.hadoop.util.*;
11.
12. public class WordCount {
13.
14.   public static class Map extends MapReduceBase implements
      Mapper<LongWritable, Text, Text, IntWritable> {
15.     private final static IntWritable one = new IntWritable(1);
16.     private Text word = new Text();
17.
18.     public void map(LongWritable key, Text value,
      OutputCollector<Text, IntWritable> output, Reporter reporter) throws
      IOException {
19.       String line = value.toString();
20.       StringTokenizer tokenizer = new StringTokenizer(line);
21.       while (tokenizer.hasMoreTokens()) {
22.         word.set(tokenizer.nextToken());
23.         output.collect(word, one);
24.       }
25.     }
26.   }
27.
```

```
28.   public static class Reduce extends MapReduceBase implements
      Reducer<Text, IntWritable, Text, IntWritable> {
29.     public void reduce(Text key, Iterator<IntWritable> values,
      OutputCollector<Text, IntWritable> output, Reporter reporter) throws
      IOException {
30.       int sum = 0;
31.       while (values.hasNext()) {
32.         sum += values.next().get();
33.       }
34.       output.collect(key, new IntWritable(sum));
35.     }
36.   }
37.
38.   public static void main(String[] args) throws Exception {
39.     JobConf conf = new JobConf(WordCount.class);
40.     conf.setJobName("wordcount");
41.
42.     conf.setOutputKeyClass(Text.class);
43.     conf.setOutputValueClass(IntWritable.class);
44.
45.     conf.setMapperClass(Map.class);
46.     conf.setCombinerClass(Reduce.class);
47.     conf.setReducerClass(Reduce.class);
48.
49.     conf.setInputFormat(TextInputFormat.class);
50.     conf.setOutputFormat(TextOutputFormat.class);
51.
52.     FileInputFormat.setInputPaths(conf, new Path(args[0]));
53.     FileOutputFormat.setOutputPath(conf, new Path(args[1]));
54.
55.     JobClient.runJob(conf);
57.   }
58. }
59.
```


https://hadoop.apache.org/docs/r1.0.4/mapred_tutorial.html#Example%3A+WordCount+v1.0

Conclusion


Scraping is hard …
… even in 2016!
https://i.ytimg.com/vi/Nt-dP2AiOkQ/maxresdefault.jpg

Berlin Buzzwords #1 Top Speaker!

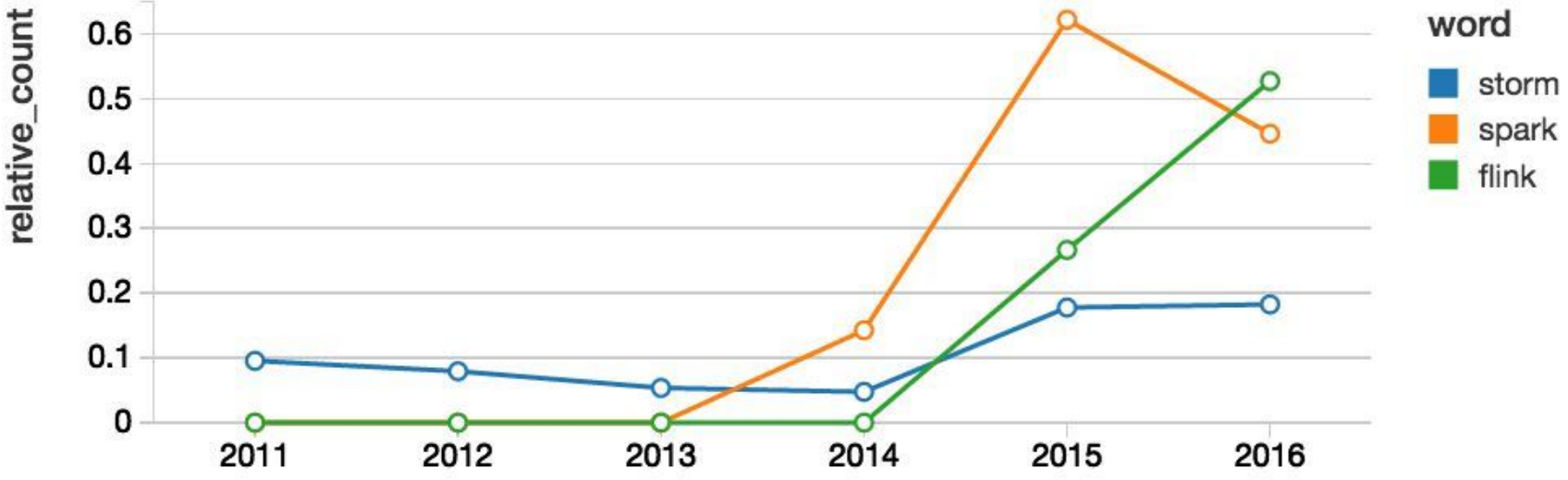
More Berlin Buzzwords Top Speakers!

Eric Evans

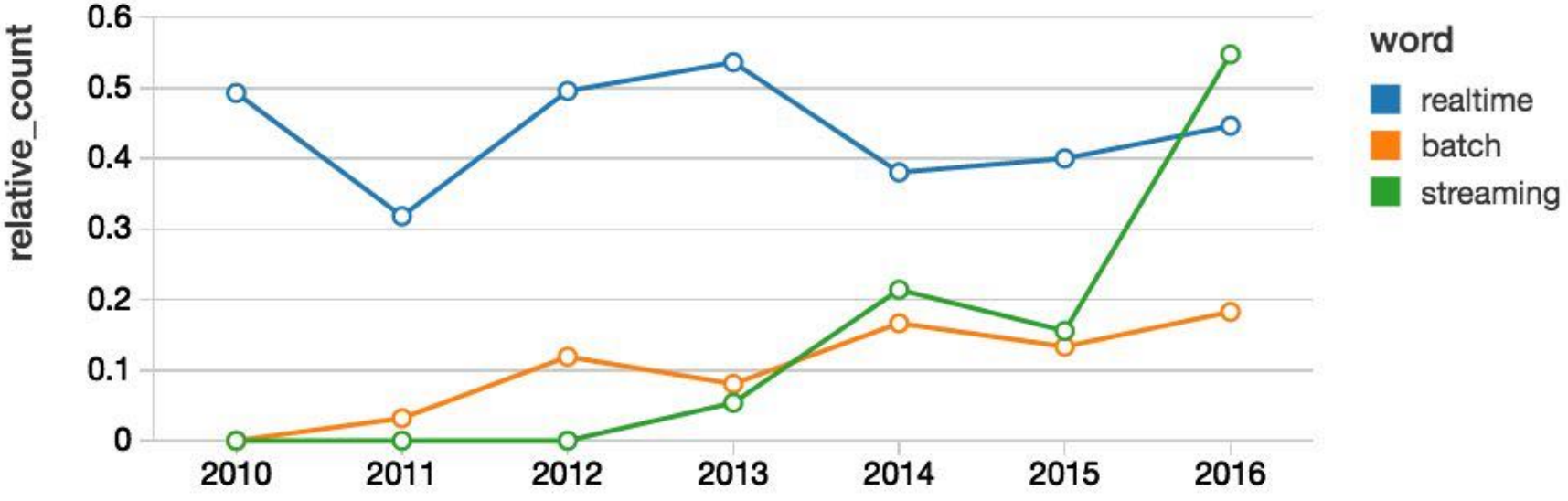
Uwe Schindler

Grant Ingersoll

Flink Took Over Spark this Year!



The Age of Streaming is There! (Who would have guessed?!?)



Thank You So Much!
Questions?

Twitter: @ctavan
https://github.com/ctavan/bbuzz2016

christoph@mbr-targeting.com
https://mbr-targeting.com