FITTING HUMANS STORIES IN LIST COLUMNS

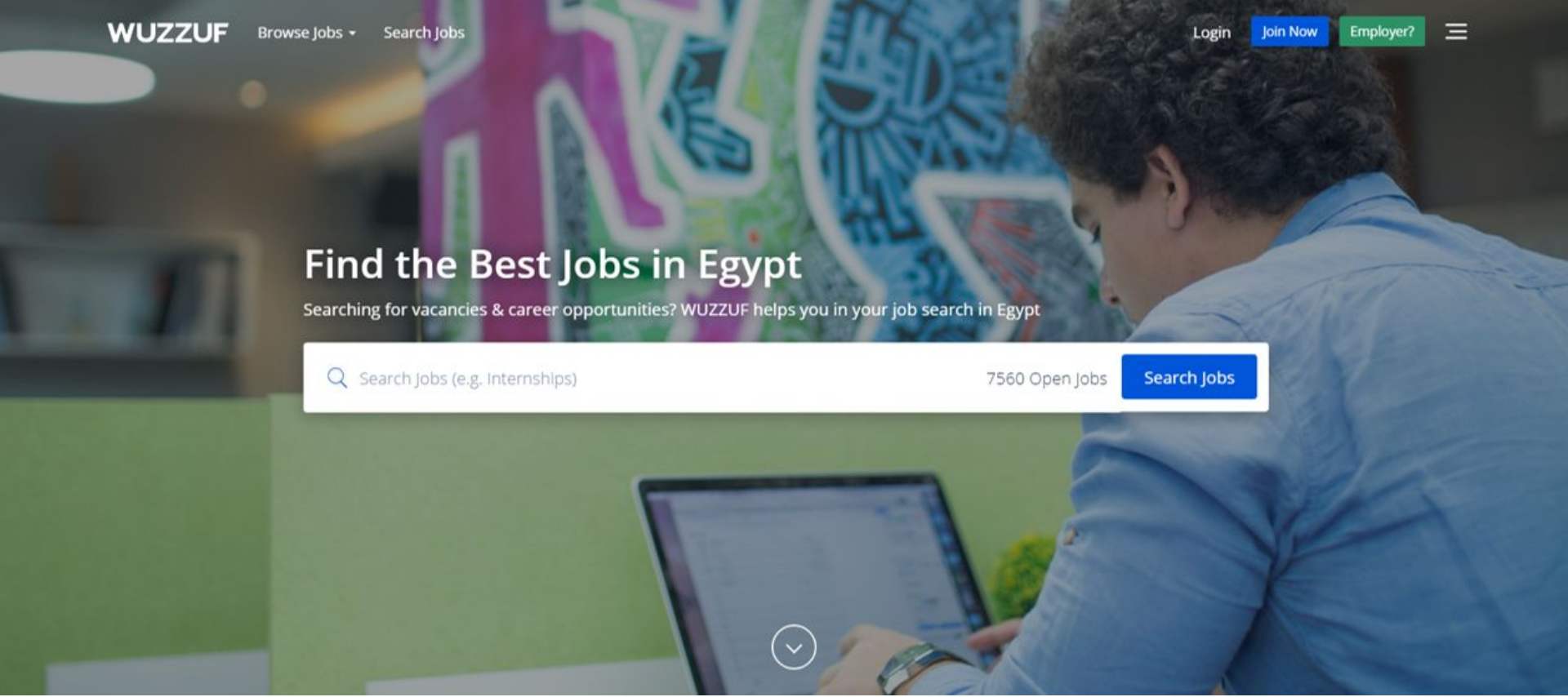
Cases from an Online Recruitment Platform

Omayma Said

@OmaymaS_

Data Scientist

WUZZUF



WUZZUF The Leading Job Site in **EGYPT**

19th Century

Adolphe Quetelet

19th Century

THE AVERAGE MAN

(L'homme Moyen)

Adolphe Quetelet

THE AVERAGE MAN

Physical

Weight, Height
(Body Mass Index)

THE AVERAGE MAN

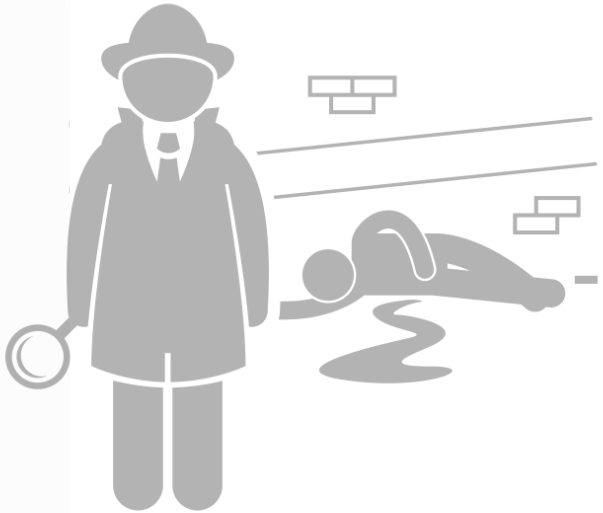
Social

Marriage

The AVERAGE MAN

Moral

Crimes



For Quetelet

THE AVERAGE MAN

=

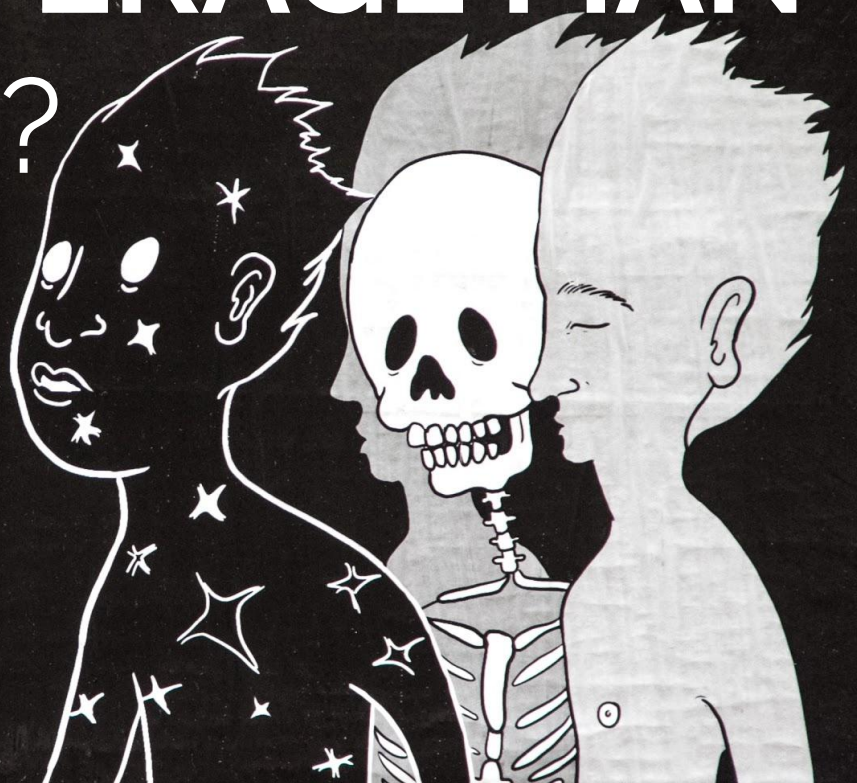
PERFECTION

“If an individual at any given epoch of society possessed all the qualities of the **AVERAGE MAN**, he would represent all that is great, good, or beautiful.”

Adolphe Quetelet

Who Is The **"AVERAGE MAN"** in Your Society?

Are You Just a Deviant from The **“AVERAGE MAN”** ?



Many **Disagree !**

Now...

Now...

Tremendous Growth of Data

Misuse of **SUMMARY STATISTICS**

Misuse of

SUMMARY STATISTICS

bon appétit

POP CULTURE

Just How Food-Obsessed Is the Typical Millennial?

Millennials are forking over $96 billion a year on food. Here, a less-than-scientific look at their purchases.

FEBRUARY 16, 2016
BY ANNA PEELE

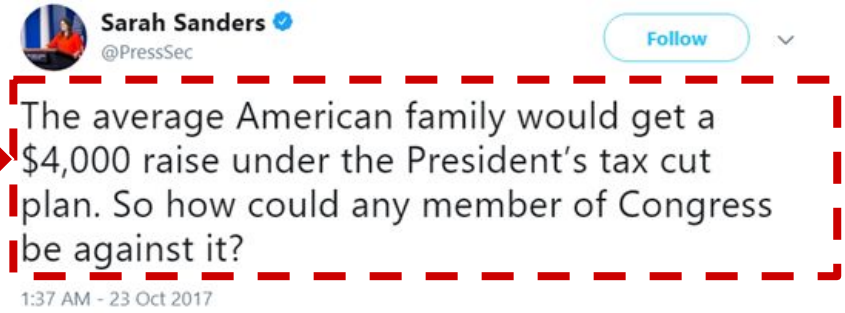
Bon Appétit
@bonappetit

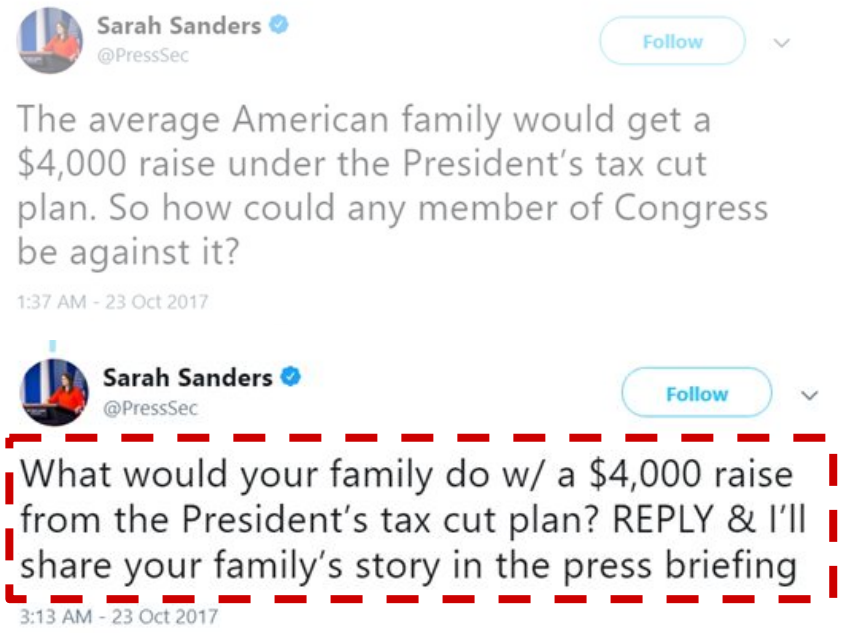
Follow

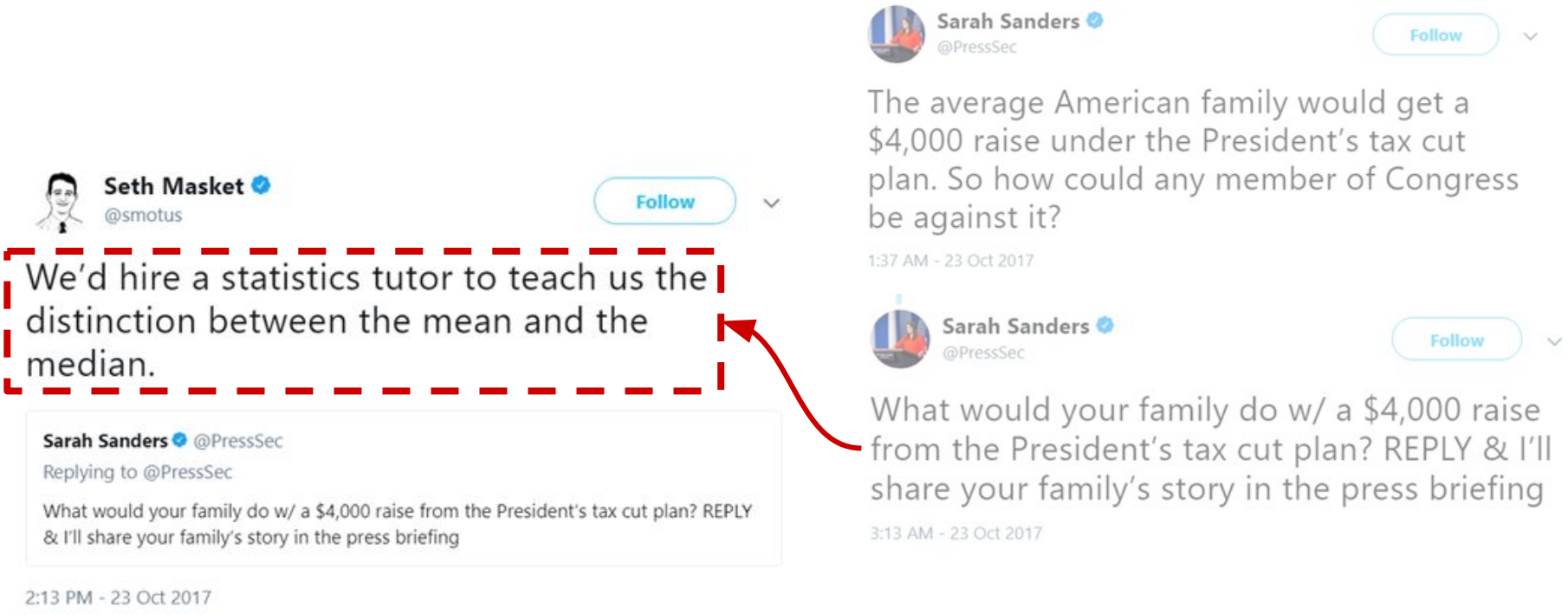
The average millennial spends $96 billion on food. Here's how we break it down
bonap.it/bxcVWz7

Misuse of

SUMMARY STATISTICS




Sarah Sanders
@PressSec
Follow
The average American family would get a $4,000 raise under the President's tax cut plan. So how could any member of Congress be against it?
1:37 AM - 23 Oct 2017
Sarah Sanders
@PressSec
Follow
What would your family do w/ a $4,000 raise from the President's tax cut plan? REPLY & I'll share your family's story in the press briefing
3:13 AM - 23 Oct 2017


Seth Masket
@smotus
Follow
We'd hire a statistics tutor to teach us the distinction between the mean and the median.
Sarah Sanders @PressSec
Replying to @PressSec
What would your family do w/ a $4,000 raise from the President's tax cut plan? REPLY & I'll share your family's story in the press briefing
2:13 PM - 23 Oct 2017
Sarah Sanders
@PressSec
Follow
The average American family would get a $4,000 raise under the President's tax cut plan. So how could any member of Congress be against it?
1:37 AM - 23 Oct 2017
Sarah Sanders
@PressSec
Follow
What would your family do w/ a $4,000 raise from the President's tax cut plan? REPLY & I'll share your family's story in the press briefing
3:13 AM - 23 Oct 2017

Joel Grus
@joelgrus

Following

there comes a time in every data scientist's career when management asks you to take an average of averages, and that's when you find out what you're really made of

7:22 PM - 28 Mar 2018 from Seattle, WA

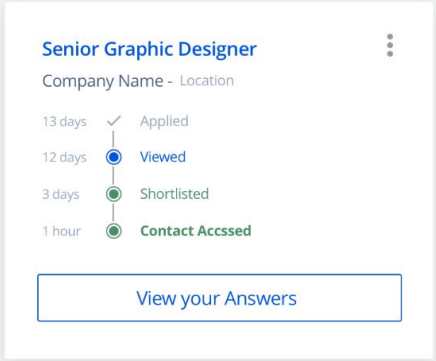
WUZZUF

Explore the Right Jobs & Career Opportunities

Explore feed knows what you need, based on your career interests, will find you what you are searching for. And don't worry about too many opportunities, you can always save them for later.

WUZZUF

Track Your Application, the Easy Way



Track your job application status whether it is viewed, shortlisted, rejected, or if a company accessed your contacts. With the tracking feature, you will be one step ahead on your job hunting plan.

WUZZUF

Take Control Over Your Exposure

With WUZZUF new profile you are in full control. You can make it public so you can use it to brand yourself, or make it visible only for employers to invite you to apply.

Get Started Now

What Do We Optimize For?

1 **Quality**

2 **Quantity**

3 **Relevance**

Matching Jobs & Job Seekers

Let’s talk about **DATA**

KPIs

METRICS

“The **average** job seeker applies for N jobs per month”

Me:

“The **average** number of applications per job this month is GREAT”

Me:

What **AVERAGE**
Do You Measure?

Who is The
AVERAGE
Job Seeker?

Can We Tell
Better **STORIES**
About Our Users?

We can tell better stories with....

Contextual Understanding + **Effective Data Analysis**

Contextual Understanding

- Culture
- Socioeconomic Status
- Market Dynamics

+

Effective Data Analysis

Contextual Understanding

+

Effective Data Analysis

- Mindset
- Workflow
- Framework/Tools

Contextual Understanding

- Culture
- Socioeconomic Status
- Market Dynamics

+

Effective Data Analysis

- Mindset
- Workflow
- Framework/Tools

Contextual Understanding

+

Effective Data Analysis

=

Better Stories

Contextual Understanding

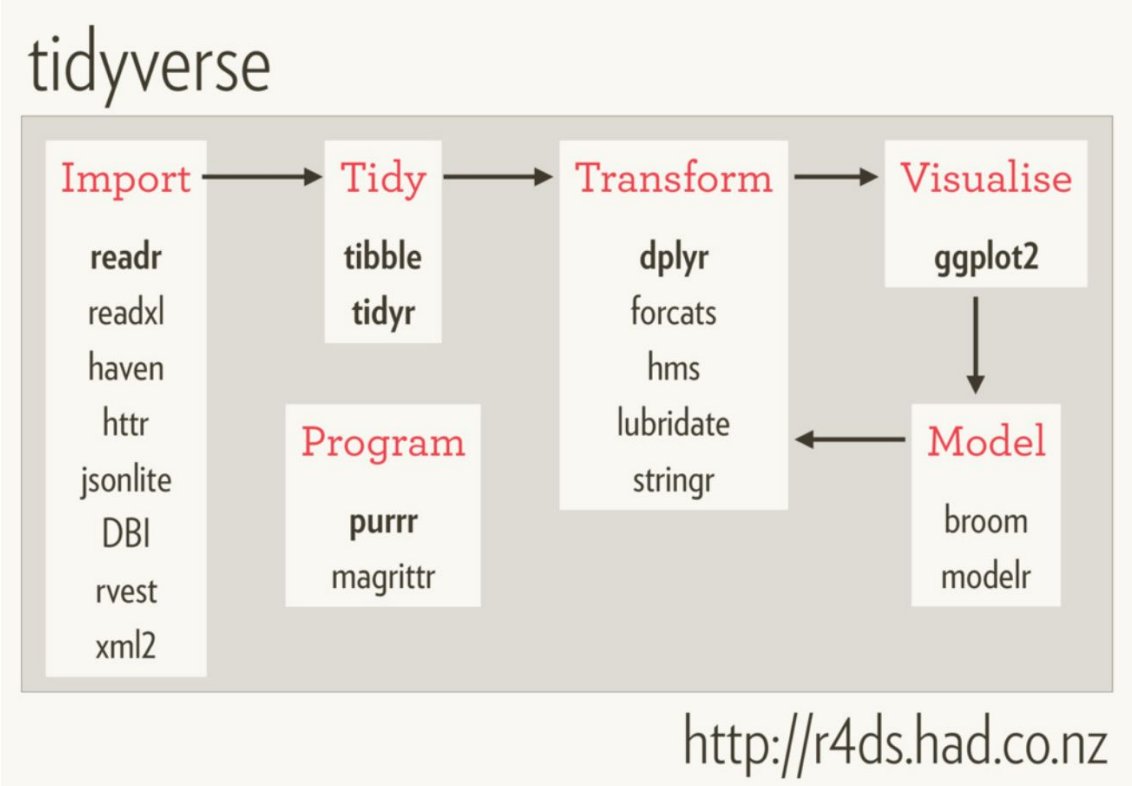
+

Effective Data Analysis

=

Actionable Insights

Framework/Tools



+ **Compatible Packages**

https://speakerdeck.com/hadley/tidyverse

The Tidyverse

Let's focus on

3 Main Concepts

Three Main Concepts

1 Tidy Data

by: @_inundata & @jcheng

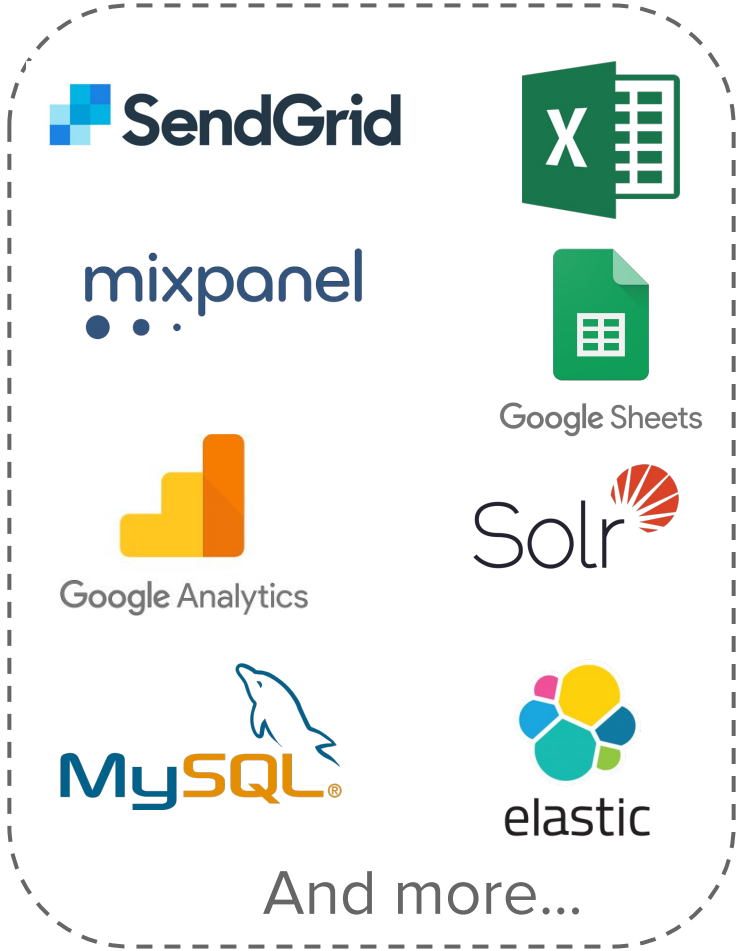
Three Main Concepts

1 Tidy Data

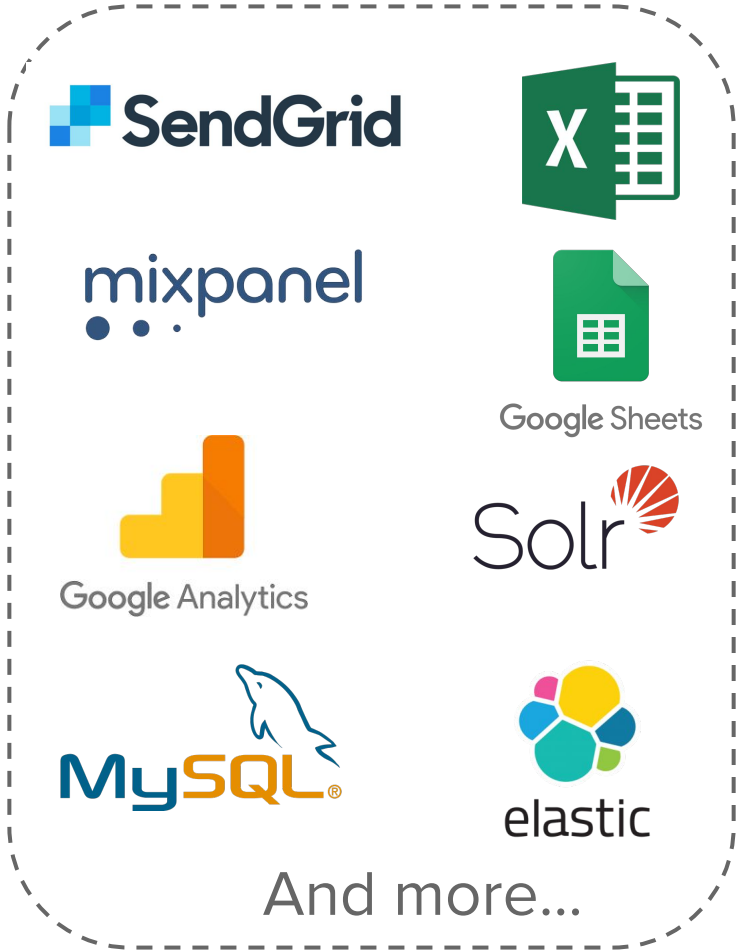
A variable in a column
An observation in a row
Tidy your data
And here you go!

[**tibble, tidyr, dplyr,** and friends]

Data comes
from **different**
SOURCES



Data comes
in **different**
FORMATS



Data comes
in **different**

FORMATS

DATAFRAME
(TIBBLE)

Tidy Data

user	job_id	job_title	company	application_date
Sara 	A1234	Software Developer	Company A	2017-01-02
Sara 	A1568	Senior Software Engineer	Company B	2017-03-02
Sara 	A1590	Software Engineer	Company C	2017-03-03
…...	…...	….	….	….
Omar 	A1234	Software Developer	Company A	2017-01-03
Omar 	A1580	Android Developer	Company C	2017-01-20
…...	….	….	….	…...

Three Main Concepts

2 Nested Data

2 Nested Data

One row per group

Instead of

One row per observation

[tidyr]

Nested Data

user	job_id	job_title	company	application_date
Sara 	A1234	Software Developer	Company A	2017-01-02
Sara 	A1568	Senior Software	Company B	2017-03-02
Sara 				
…...				
Omar 				
Omar 				
….				

```
user_data %>%
    group_by(user) %>%
    nest(.key = "applications")
```

user	applications
Sara 	<Tibble [3 x 4]>
Omar 	<Tibble [2 x 4]>
….	…...

Nested Data

user	job_id	job_title	company	application_date
Sara 	A1234	Software Developer	Company A	2017-01-02
Sara 	A1568	Senior Software	Company B	2017-03-02
Sara 				
……				
Omar 				
Omar 				
……				

```
job_data %>%
    group_by(job_id) %>%
    nest(.key = "applications")
```

job_id	applications
A1234 	<Tibble [2 x 4]>
A1568 	<Tibble [30 x 4]>
A1590 	<Tibble [100 x 4]>
A1580 	<Tibble [120 x 4]>

3 Functional Programming

3 Functional Programming

Handle iteration problems powerfully and emphasize the actions rather than the objects

[purrr]

Let's store models in columns

job_id	applications	app_count
A5638 	<tibble [362 x 27]>	362
A8957 	<tibble [110 x 27]>	110
.....	.....	.....

```
job_app_data<- job_app_data %>%
  mutate(glm_model = map(app_data, ~ glm(viewed ~ app_day,
                                         data = .x, family = binomial)))
```

Let's store models in columns

job_id	applications	app_count	**glm_model**
A5638 	<tibble [362 x 27]>	362	**<S3: glm>**
A8957 	<tibble [110 x 27]>	110	**<S3: glm>**
.....	.....	.....	....

```
job_app_data<- job_app_data %>%
  mutate(glm_model = map(app_data, ~ glm(viewed ~ app_day,
                                         data = .x, family = binomial)))
```

Iterate and answer more questions

user	applications	preferences
Sara 	<tibble [2 x 10]>	<tibble [4 x 10]>
Omar 	<tibble [2 x 15]>	<tibble [2 x 10]>
…..	…..	….

```
user_data <- user_data %>%
  mutate(common_jobs = map2(applications, preferences,
                            ~intersect(.x[["job_title"],.y[["job_title"]])
```

Iterate and answer more questions

user	applications	preferences	**common_jobs**
Sara 	<tibble [2 x 10]>	<tibble [4 x 10]>	**<chr [2]>**
Omar 	<tibble [2 x 15]>	<tibble [2 x 10]>	**<chr [0]>**
…..	…..	….	

```
user_data <- user_data %>%
  mutate(common_jobs = map2(applications, preferences,
                            ~intersect(.x[["job_title"],.y[["job_title"]])
```

Let’s Look Closer !

Problem

Overall growth and **good KPIs**

Shortage in applications for certain **Software Development** jobs

Problem

Shortage in applications for certain **Software Development** jobs

---

Dissatisfied Employers

Problem

Shortage in applications for certain
Software Development jobs

Flagged by different sources

Problem

Shortage in applications for certain
Software Development jobs

Masked by high-level metrics


The game,

Hypotheses

1 Talent Shortage

What if we just have a small pool of job seekers who are interested in the affected jobs?

Hypotheses

2 Irrelevant Jobs

Maybe employers are not catching up with the global trends or job seekers aspirations!

Hypotheses

3

Hidden Jobs

What if some jobs do not get enough exposure in the search/recommendation pages?

Investigation

1st The **Job's** Side

The Job's Side

What about applications details per job?

```
job_app <- left_join(jobs, apps,
                     by = c("job_id" = "job_id")) %>%
  group_by(job_id, job_title, post_date) %>%
  nest(.key = "app_data")
```

```
# A tibble: 2,934 x 5
    job_id                     job_title           app_data app_count
     <chr>                         <chr>             <list>     <int>
1 5e934219 Junior Communication Engineer <tibble [219 x 4]>       219
2 cba698f2                Web Developer  <tibble [26 x 4]>        26
3 60596486               Office Manager  <tibble [45 x 4]>        45
4 f4343410   Real Estate Sales Executive  <tibble [29 x 4]>        29
5 124aae63   Senior SharePoint Developer  <tibble [17 x 4]>        17
# ... with 2,929 more rows, and 1 more variables: post_date <date>
```

The Job's Side

Applications Details

```
# A tibble: 219 x 4
  application_id application_date  user_id app_day
           <chr>           <date>    <chr>  <time>
1       66851a93       2017-04-03 8d6cfddf  0 days
2       c71e39f5       2017-04-03 c6223d74  0 days
3       e53333f3       2017-04-03 56c5c8df  0 days
# ... with 216 more rows
```

```
# A tibble: 2,934 x 5
    job_id                     job_title           app_data app_count
     <chr>                         <chr>             <list>     <int>
1 5e934219 Junior Communication Engineer <tibble [219 x 4]>       219
2 cba698f2                 Web Developer  <tibble [26 x 4]>        26
3 60596486                Office Manager  <tibble [45 x 4]>        45
4 f4343410   Real Estate Sales Executive  <tibble [29 x 4]>        29
5 124aae63   Senior SharePoint Developer  <tibble [17 x 4]>        17
# ... with 2,929 more rows, and 1 more variables: post_date <date>
```

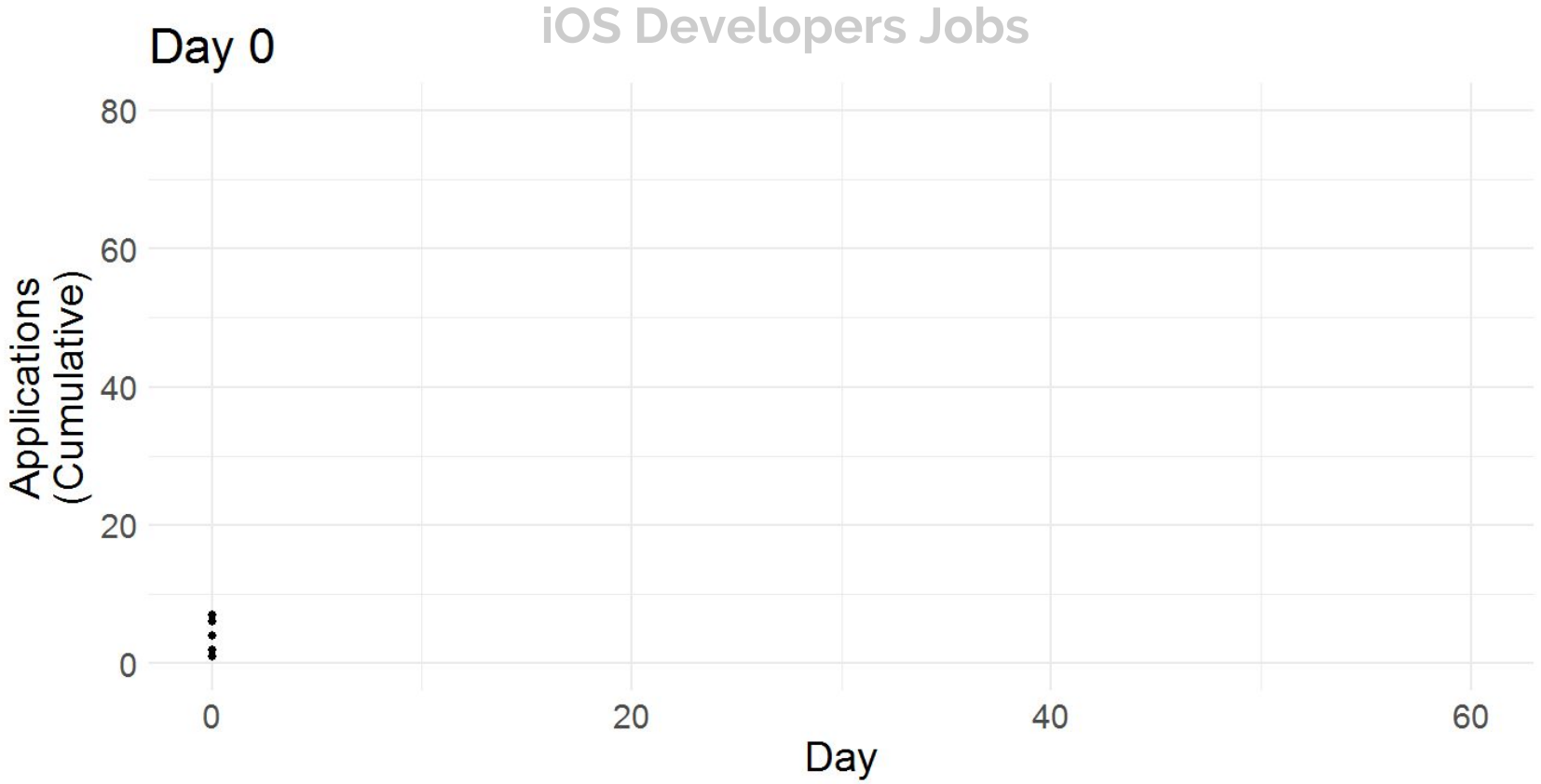
The Job's Side

What about iOS job applications?

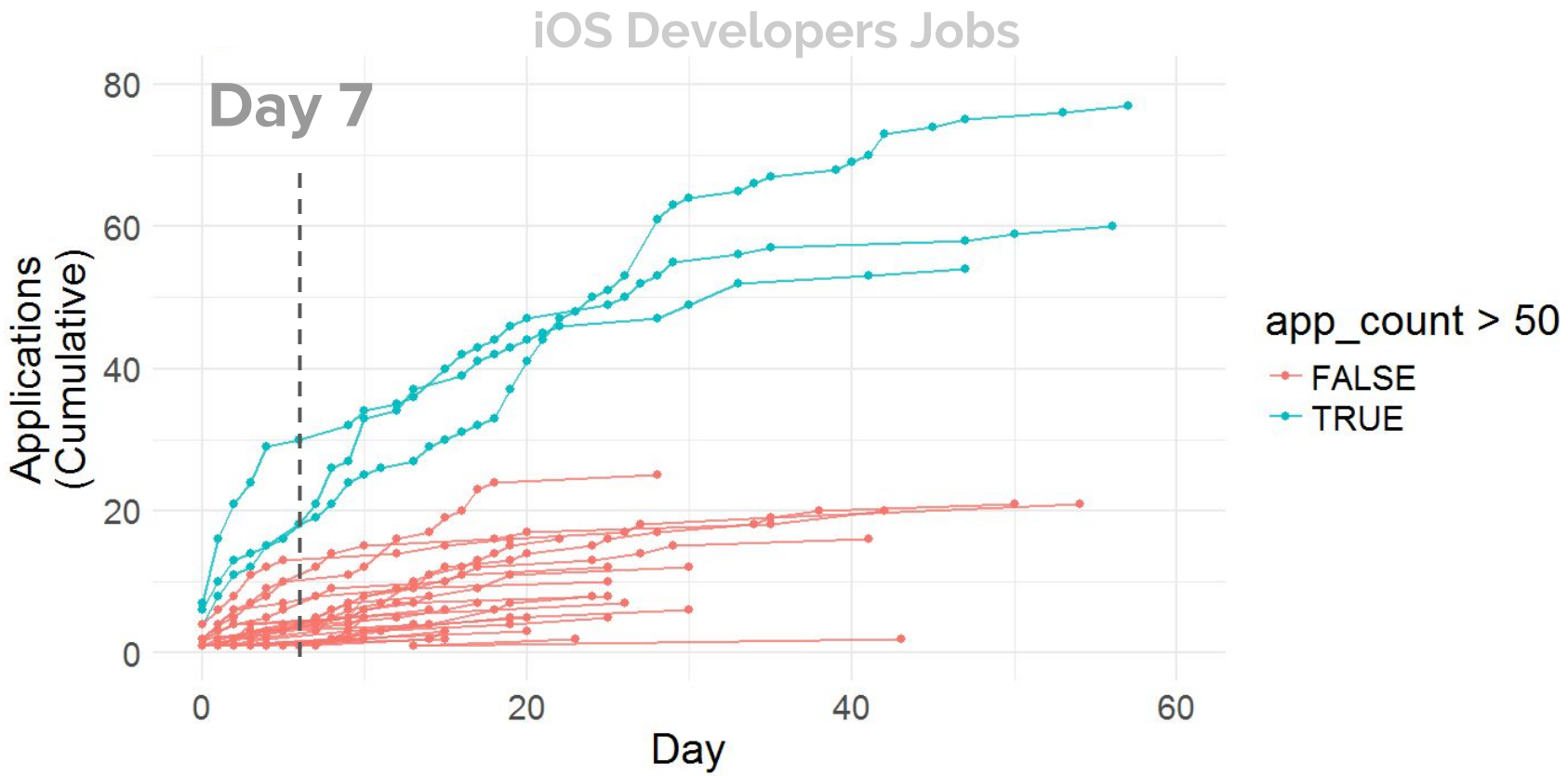
```
job_app_ios <- job_app %>%
  filter(grepl("\\biOS\\b", job_title ))
```

```
# A tibble: 34 x 4
    job_id     job_title          app_data app_count
     <chr>         <chr>            <list>     <int>
1 54344870 iOS Developer  <tibble [2 x 4]>         2
2 d647f642 iOS Developer  <tibble [2 x 4]>         2
3 b3e9f878 iOS Developer  <tibble [6 x 4]>         6
4 b137842c iOS Developer  <tibble [7 x 4]>         7
5 7b1f1998 iOS Developer <tibble [10 x 4]>        10
# ... with 29 more rows
```

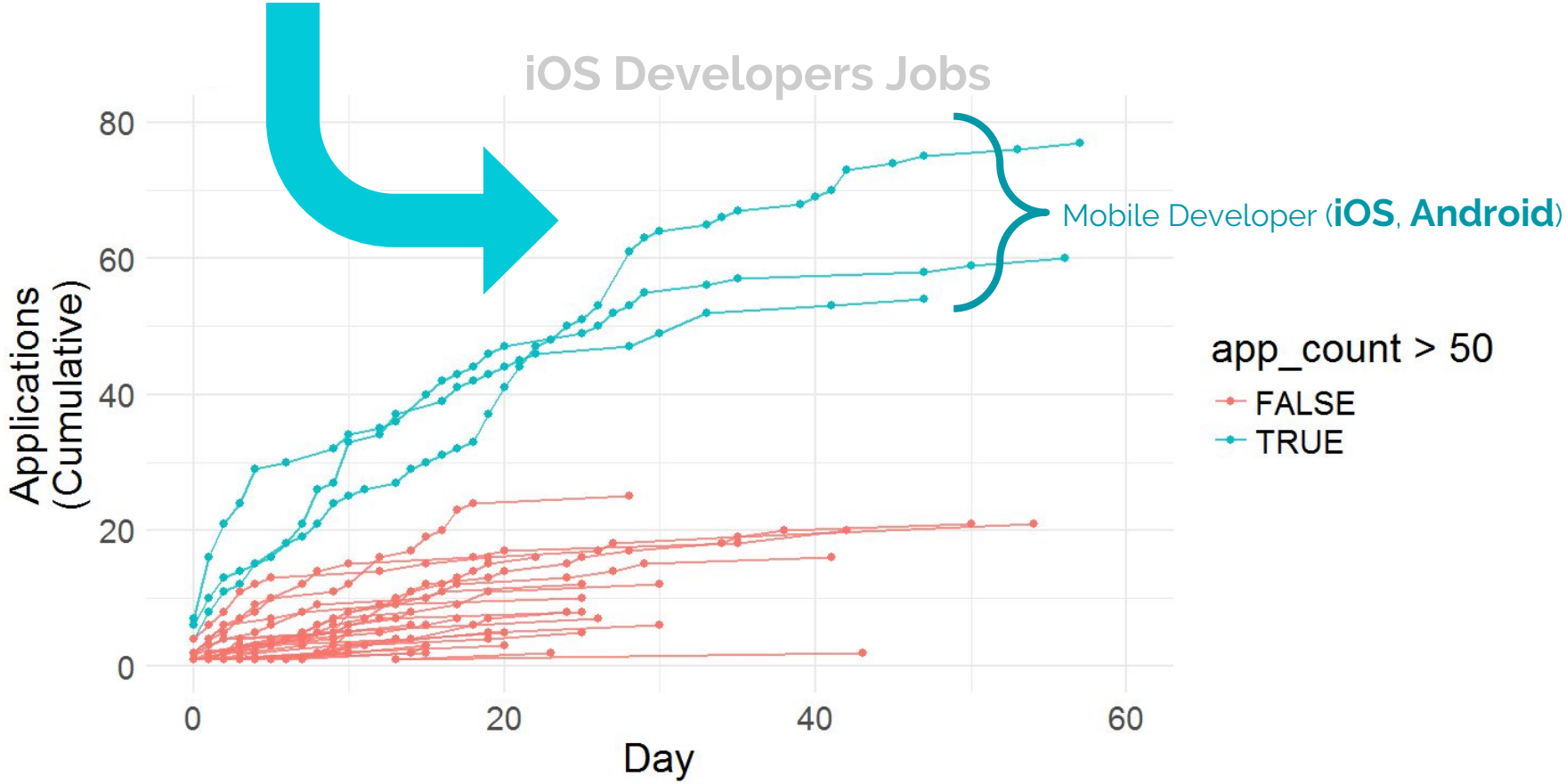
Job Applications Growth over time

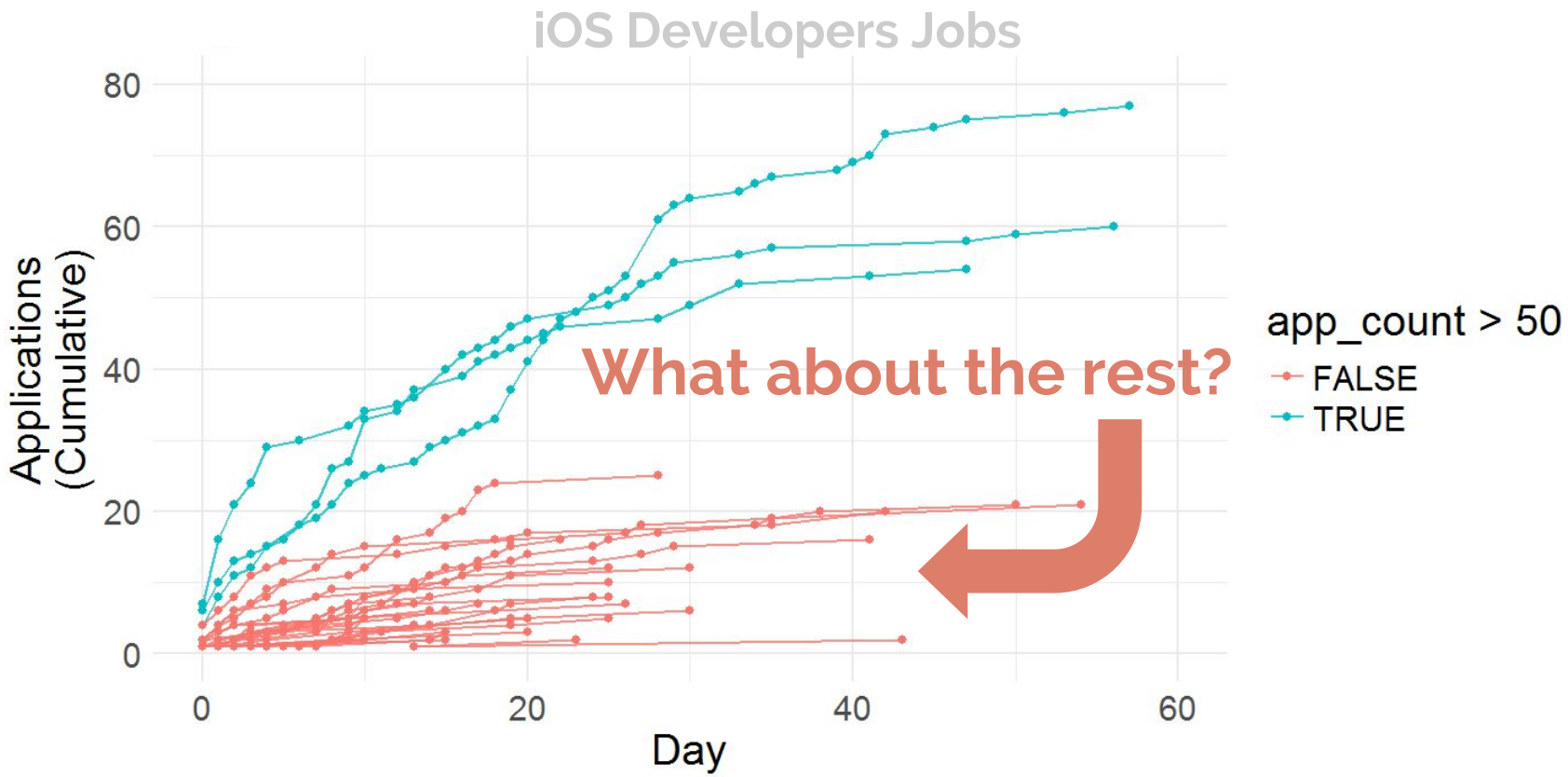


What happens to job posts on day X?

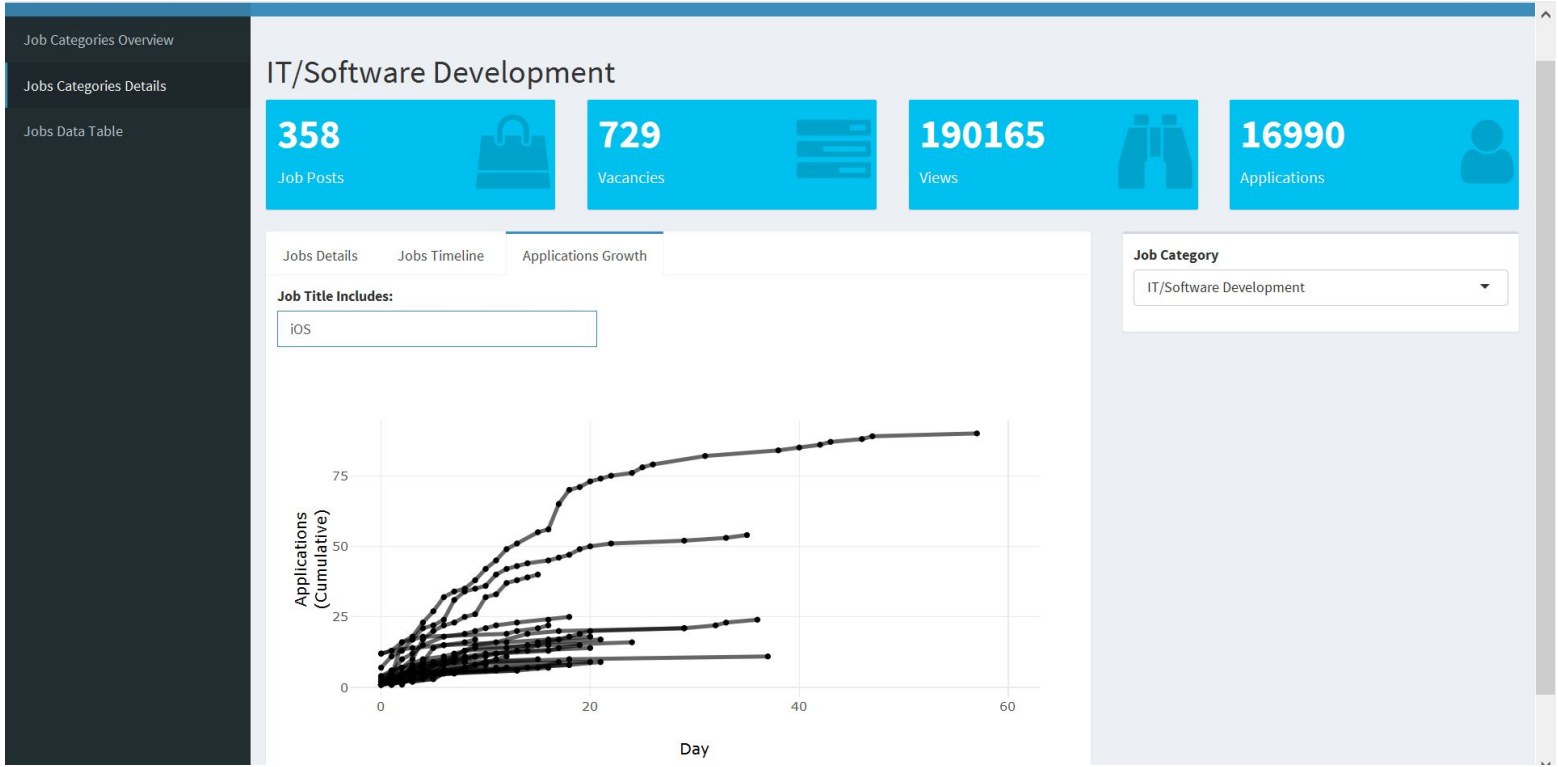


What is special about these jobs?




iOS Developers Jobs
Applications (Cumulative)
80
60
40
20
0
0
20
40
60
Day
What about the rest?
app_count > 50
FALSE
TRUE

More with Shiny...



*Sample of Wuzzuf Job Posts

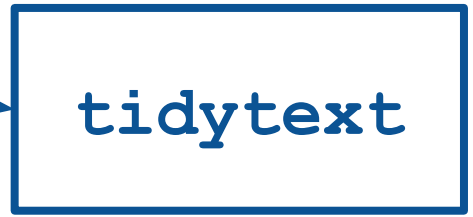
Investigation

2nd The **Job Seeker's** Side

The Job Seeker's Side

How do job seekers fill their profiles?

```
js_data_details <- js_data%>%
  filter(grepl("\\bios\\b", query_list)) %>%
  mutate(kw_freq = map(query_list, ~ tidytext::unnest_tokens(.x, term, value,
                                                             token = "regex",
                                                             pattern = "|") %>%
                         count(term, sort = TRUE)))
```



The Job Seeker's Side

How do job seekers fill their profiles?

```
js_data_details <- js_data%>%
  filter(grepl("\\bios\\b", query_list)) %>%
  mutate(kw_freq = map(query_list, ~ tidytext::unnest_tokens(.x, term, value,
                                                             token = "regex"
                                                             pattern =
                       count(term, sort = TRUE)))
```

```
# A tibble: 388 x 3
   user_id query_list          kw_freq
     <chr>     <list>            <list>
1 4003e037  <chr [1]> <tibble [19 x 2]>
2 9d0ba246  <chr [1]> <tibble [20 x 2]>
3 eeac5b9e  <chr [1]> <tibble [24 x 2]>
4 32a1e586  <chr [1]> <tibble [22 x 2]>
5 f48c2ee0  <chr [1]> <tibble [15 x 2]>
# ... with 383 more rows
```

Job Seeker's Keywords

```
# A tibble: 22 x 2
                term     n
               <chr> <int>
1          asp net       3
2  android engineer      1
3          android       1
4              asp       1
5               c#       1
# ... with 17 more rows
```

The Job Seeker's Side

What about the repetition in the extracted keywords?

```
js_data_details <- js_data %>%
  filter(grepl("\\bios\\b", query_list)) %>%
  mutate(kw_freq = map(query_list, query_kw_freq)) %>%
  mutate(kw_count = map_int(kw_freq, nrow)) %>%
  mutate(kw_freq_max = map_int(kw_freq, ~max(.x[["freq"]])))
```

The Job Seeker's Side

What about the repetition in the extracted keywords?

```
js_data_details <- js_data %>%
  filter(grepl("\\bios\\b", query_list)) %>%
  mutate(kw_freq = map(query_list, query_kw_freq)) %>%
  mutate(kw_count = map_int(kw_freq, nrow)) %>%
  mutate(kw_freq_max = map_int(kw_freq, ~max(.x[["freq"]])))
```

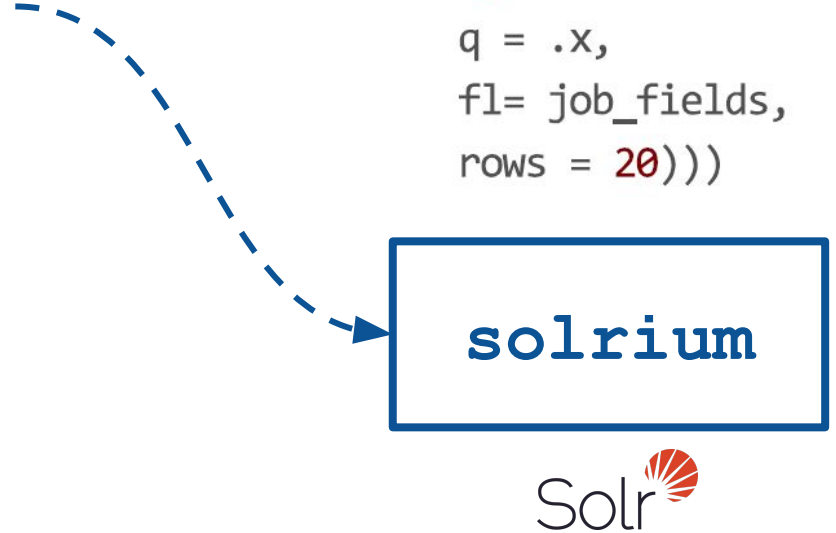
Summaries from Job Seeker's Keywords

```
# A tibble: 388 x 5
   user_id query_list          kw_freq          kw_count kw_freq_max
     <chr>     <list>            <list>             <int>       <int>
1 4003e037  <chr [1]> <tibble [19 x 2]>>              19           2
2 9d0ba246  <chr [1]> <tibble [20 x 2]>>              20           5
3 eeac5b9e  <chr [1]> <tibble [24 x 2]>>              24           3
4 32a1e586  <chr [1]> <tibble [22 x 2]>>              22           3
5 f48c2ee0  <chr [1]> <tibble [15 x 2]>>              15           5
# ... with 383 more rows
```

The Job Seeker's Side

Which jobs match each user's profile?

```
js_data_details <- js_data_details %>%
  mutate(jobs_search_results = map(query_list,
                                   ~ solrium::solr_search("jobs",
                                                          q = .x,
                                                          fl= job_fields,
                                                          rows = 20)))
```



The Job Seeker's Side

Which jobs match each user's profile?

```
js_data_details <- js_data_details %>%
  mutate(jobs_search_results = map(query_list,
                                   ~ solrium::solr_search("jobs",
                                                          q = .x,
                                                          fl= job_fields,
                                                          rows = 20)))
```

```
# A tibble: 388 x 6
   user_id query_list          kw_freq jobs_search_results kw_count
     <chr>     <list>            <list>              <list>    <int>
1 4003e037  <chr [1]> <tibble [19 x 2]>   <tibble [20 x 5]>       19
2 9d0ba246  <chr [1]> <tibble [20 x 2]>   <tibble [20 x 5]>       20
3 eeac5b9e  <chr [1]> <tibble [24 x 2]>   <tibble [20 x 5]>       24
4 32a1e586  <chr [1]> <tibble [22 x 2]>   <tibble [20 x 5]>       22
5 f48c2ee0  <chr [1]> <tibble [15 x 2]>   <tibble [20 x 5]>       15
# ... with 383 more rows, and 1 more variables: kw_freq_max <int>
```

The Job Seeker's Side

Which jobs match each user's profile?

Recommended Jobs Details

```
# A tibble: 20 x 5
    job_id                          job_title            post_date
     <chr>                              <chr>                <chr>
1 4a871cd4 Senior Web & Mobile Apps Developer 2017-03-04T18:03:01Z
2 48cd2159              Mobile Apps Developer 2017-06-20T00:00:00Z
3 4ec0abe3             Full Stack Team Leader 2017-02-21T09:49:57Z
4 694443c0            .NET Software Developer 2017-03-07T16:03:09Z
5 cc8381d8           Senior Android Engineer 2017-03-12T16:36:18Z
# ... with 15 more rows, and 2 more variables: max_salary <int>,
#   skills <chr>
```

```
# A tibble: 388 x 6
   user_id query_list          kw_freq jobs_search_results kw_count
     <chr>     <list>            <list>              <list>    <int>
1 4003e037  <chr [1]> <tibble [19 x 2]>  <tibble [20 x 5]>       19
2 9d0ba246  <chr [1]> <tibble [20 x 2]>  <tibble [20 x 5]>       20
3 eeac5b9e  <chr [1]> <tibble [24 x 2]>  <tibble [20 x 5]>       24
4 32a1e586  <chr [1]> <tibble [22 x 2]>  <tibble [20 x 5]>       22
5 f48c2ee0  <chr [1]> <tibble [15 x 2]>  <tibble [20 x 5]>       15
# ... with 383 more rows, and 1 more variables: kw_freq_max <int>
```


What **ACTIONS**
Did This Analysis
Trigger?

Recommended Actions

1 Talent Shortage

- Acquire more senior developers
- Activate the existing developers
- Support the community

Recommended Actions

2 Irrelevant Jobs

- Advise employers about the market
- Revisit preference-based matching

Recommended Actions

3 Hidden Jobs

- Revisit text fields indexing
- Tune field weights for scoring
- Improve mail recommendation

3 Main Concepts

Tidy Data
Nested Data
Functional Programming

Contextual Understanding + Effective Data Analysis = **Actionable Insights**

FITTING HUMANS STORIES IN LIST COLUMNS

Cases from an Online Recruitment Platform

Omayma Said

@OmaymaS_

Data Scientist

WUZZUF