SciencesPo
MÉDIALAB

CSV, Rinse, Repeat

Javascript Data Exploration

Mathieu Jacomy
Sciences Po Paris médialab
Equipex DIME-SHS

Paris Sciences Po médialab

A hybrid laboratory for social sciences:

- Researchers
- Engineers
- Designers

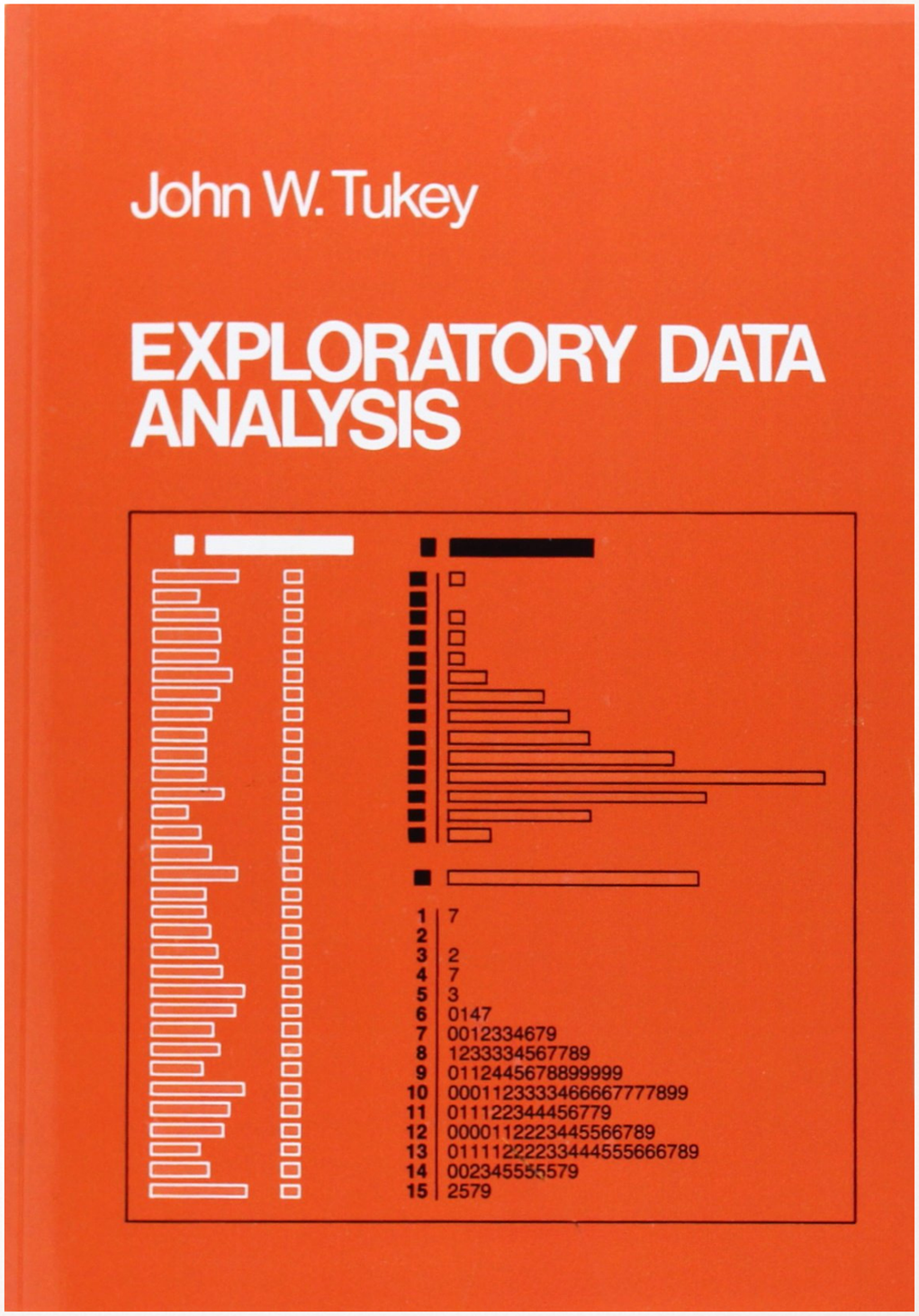
I'm a sort of « social data scientist »

Exploring data

...is not about statistical metrics

The greatest value of a picture is when it forces us to notice what we never expected to see.
— John W. Tukey

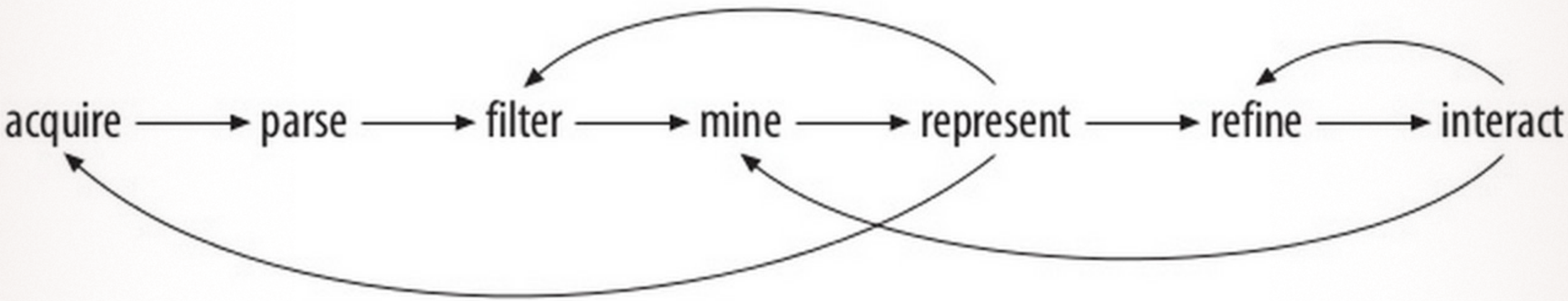
Far better an approximate answer to the right question, which is often vague, than an exact answer to the wrong question, which can always be made precise.
—John W. Tukey



What is the
right question?

Exploring data

...is about iterating facets of the data



CSV Problem #1:
Painful coding

SciencesPo
MÉDIALAB

In a spreadsheet environment,
eg. Libre Office, Excel, Open Refine...
coding features are designed for non-coders.

Consequence for non-coders:
You invest your time in nonstandard, broken languages
(you lose your time + it is still complicated)

Consequence for coders:
Editing and filtering the data is painful
(ranging from inefficient to WTF)

Painful coding, simple example

SciencesPo MÉDIALAB

Goal: retrieve years in movie titles

Rank	Rating	Movie
1	9.3/10	The Shawshank Redemption (1994)
2	9.2/10	The Godfather (1972)
3	9.0/10	The Godfather: Part II (1974)
4	9.0/10	The Dark Knight (2008)
5	8.9/10	12 Angry Men (1957)
6	8.9/10	Pulp Fiction (1994)
7	8.9/10	Schindler's List (1993)
8	8.9/10	The Lord of the Rings: The Return of the King (2003)
9	8.9/10	The Good, the Bad and the Ugly (1966)
10	8.9/10	Fight Club (1999)
11	8.8/10	The Lord of the Rings: The Fellowship of the Ring (2001)
12	8.8/10	Inception (2010)
13	8.8/10	Star Wars: Episode V - The Empire Strikes Back (1980)

Issue:
the year is coded
within the title

Simple parsing with Libre Office

SciencesPo MÉDIALAB

1. Find the year position

STXT =CHERCHE("\([0-9]{4}\)";C2;0)

	A	B	C		E	F
1	Rank	Rating	Movie			
2	1	9.3/10	The Shawshank Redemption (1994)	=CHERCHE("\([0-9]{4}\)";C2;0)	1994	
3	2	9.2/10	The Godfather (1972)	18	1972	
4	3	9.0/10	The Godfather: Part II (1974)	27	1974	

Look, it's coded in French!

2. Retrieve the string

STXT =STXT(C2;D2+1;4)

	A	B	C	D	E	F
1	Rank	Rating	Movie			
2	1	9.3/10	The Shawshank Redemption (1994)	29	=STXT(C2;D2+1;4)	
3	2	9.2/10	The Godfather (1972)	18	1972	
4	3	9.0/10	The Godfather: Part II (1974)	27	1974	

In this situation the GUI is a problem, not a solution

Simple parsing with Javascript

SciencesPo MÉDIALAB

```
// Extract year data
item.Year = item.Movie.match(/.*\(([0-9]{4})\)/)[1];
```

A real coding language is more efficient
...if you can bring your CSV in the right coding environment

CSV Problem #2:
The filter/vis gap

SciencesPo
MÉDIALAB

Code
(Filter + Edit)

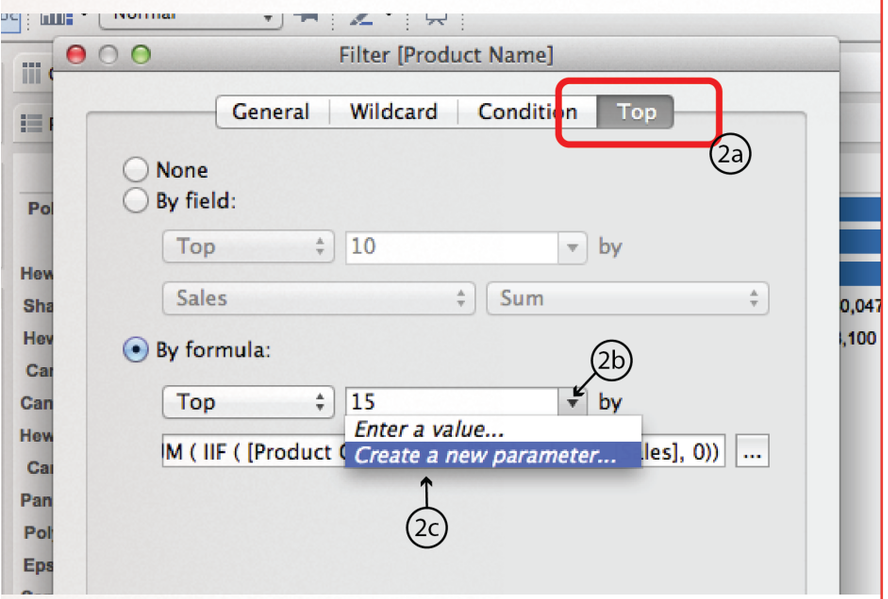
GAP

Visualize

Simple filtering in Tableau Public

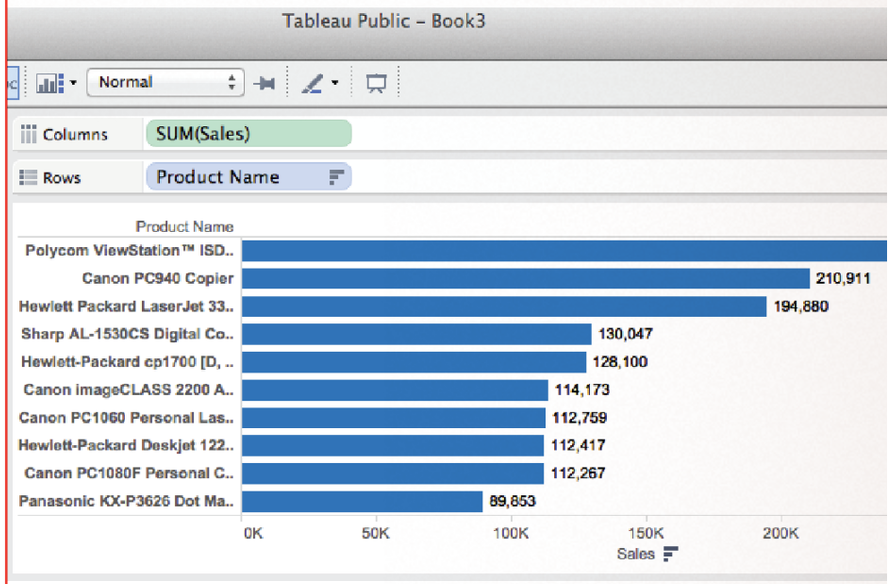
SciencesPo
MÉDIALAB

Tableau tutorial by Anne Stevens
http://stevensanne.com/tableau-tutorial-3-filters-and-parameters/



Edit the formula and/or the settings, apply and close

The modal hides the visualization



Formula + settings,
in a form,
inside a tab,
of a modal
that you open by a drag-and-drop
(true story)

Reopen the modal, select the tab, select the field

Iterating is painful

Visualization

Note:
in Libre Office or Excel, it's even worse

SciencesPo
MÉDIALAB

CODE YOUR FILTER

```
output = input.map(function(item, i){

    // Extract year data
    item.Year = item.Movie.match(/.*\(([0-9]{4})\)/)[1];

    // Clean title data (remove the year)
    item.Movie = item.Movie.replace(' ('+item.Year+')', '');

    // Clean rating
    item.Rating = item.Rating.replace('/10', '')

    return item;

}).filter(function(item, i){

    // Only the movies with a number in their title
    return item.Movie.search(/[0-9]/gi) >= 0;

});

// Hit CTRL + ENTER to run the code
```

Real Javascript coding

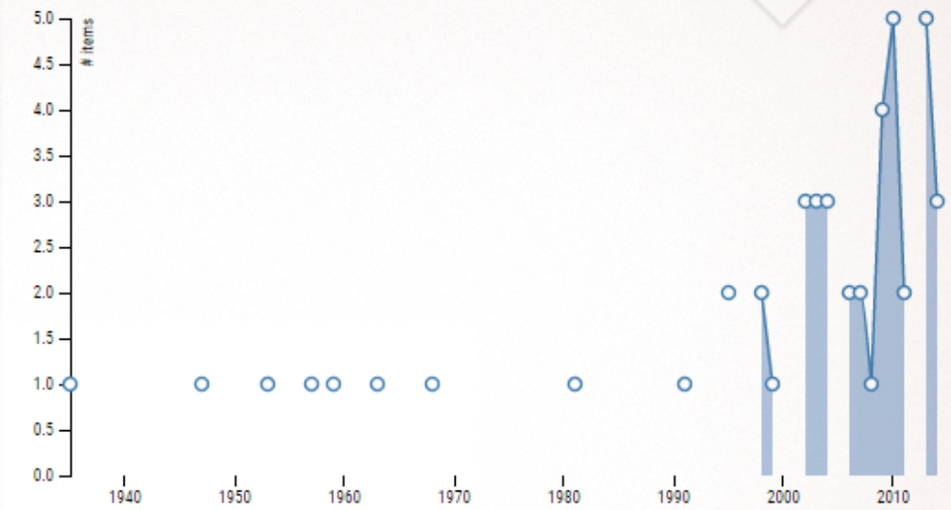
CSV, Rinse Repeat

is about **shortening the gap** between code and visualization

to foster **iterative exploration**

VISUALIZE RESULT

YEAR - YEARLY VOLUME



MOVIE - ITEMS TOP 50

1. Miracle on 34th Street (1)
2. 12 Angry Men (1)
3. Terminator 2: Judgment Day (1)
4. 3 Idiots (1)
5. Toy Story 3 (1)
6. 2001: A Space Odyssey (1)
7. The 400 Blows (1)
8. The Legend of 1900 (1)
9. Kill Bill: Vol. 1 (1)
10. 8½ (1)
11. 12 Years a Slave (1)

the Deathly Hallows: Part 2 (1)
13. 3-Iron (1)
14. Stalag 17 (1)
15. The Raid 2 (1)
16. Short Term 12 (1)
17. Kill Bill: Vol. 2 (1)
18. Special 26 (1)
19. District 9 (1)
20. 4 Months, 3 Weeks and 2 Days (1)
21. Toy Story 2 (1)
22. How

24. Se7en (1)
25. Big Hero 6 (1)
26. The 39 Steps (1)
27. (500) Days of Summer (1)
28. 3:10 to Yuma (1)
29. 300 (1)
30. 21 Grams (1)
31. 50/50 (1)
32. Harry Potter and the Deathly Hallows: Part 1 (1)
33. 25th Hour (1)
Cell 211 (1)
2: The

37. United 93 (1)
38. 28 Days Later... (1)
39. Ghost in the Shell 2: Innocence (1)
40. Apollo 13 (1)
41. 13 Assassins (1)
42. Ip Man 2 (1)
43. Mesrine Part 1: Killer Instinct (1)
44. 42(2013) (1)
45. Despicable Me 2 (1)
46. X2 (1)
47. Buffalo '66 (1)

+ ADD VIZ

Simple, ready-made visualizations

CSV, Rinse, Repeat

...is a proposition to solve these problems during exploration

A simple accessible tool
- Single web page

A Javascript coding environment
- A standard coding panel
- CSV Import + Export
- Basic preview

A layout designed to get rid of the filter/vis gap
- Input + code on the left
- Output + visualization on the right

Demo time!

http://tools.medialab.sciences-po.fr/csv-rinse-repeat/

CSV, Rinse, Repeat

Sample datasets:

- Movies
- NASA lab facilities
- OECD BEPS consultation actors

+ more tools

SciencesPo
MÉDIALAB

LOAD CSV

CLICK or DRAG A FILE

INPUT
152610 r

PREVIEW
3

	created_at	from_user_name	text	filter
937.0	2016-04-24T08:18:57	Jeremiahjones88	LIVE on #Periscope: Sonnet Sunday #shakespeare's bday https://www.periscope.tv/w/aesncDFIV0t5TnpMTVZqQWd8MUJkR1IEVIZYem9LWG0EqIIRES2vMPIQg5HxWYX6we_6DVazsiFQ7Ov06IqY	
382.0	2016-04-12T16:26:22	Lynchypoos	RT @britishlibrary: Here's a sneak peek at our exhibition, opening this Friday. What's your favourite #Shakespeare	

CODE YOUR FILTER

```
output = input.filter(function(item, i){

    if (
          // Check that latitude and longitude are valid
          item.lat && item.lng
          && !isNaN(+item.lat) && item.lat >= -90 && item.lat <
          && !isNaN(+item.lng) && item.lng >= -180 && item.lng
    ) {
        // If they are, concatenate them for the vis module
        item.coordinates = item.lat + '::' + item.lng
        return true
    } else {
        // If not, filter them
        return false
    }

});
```

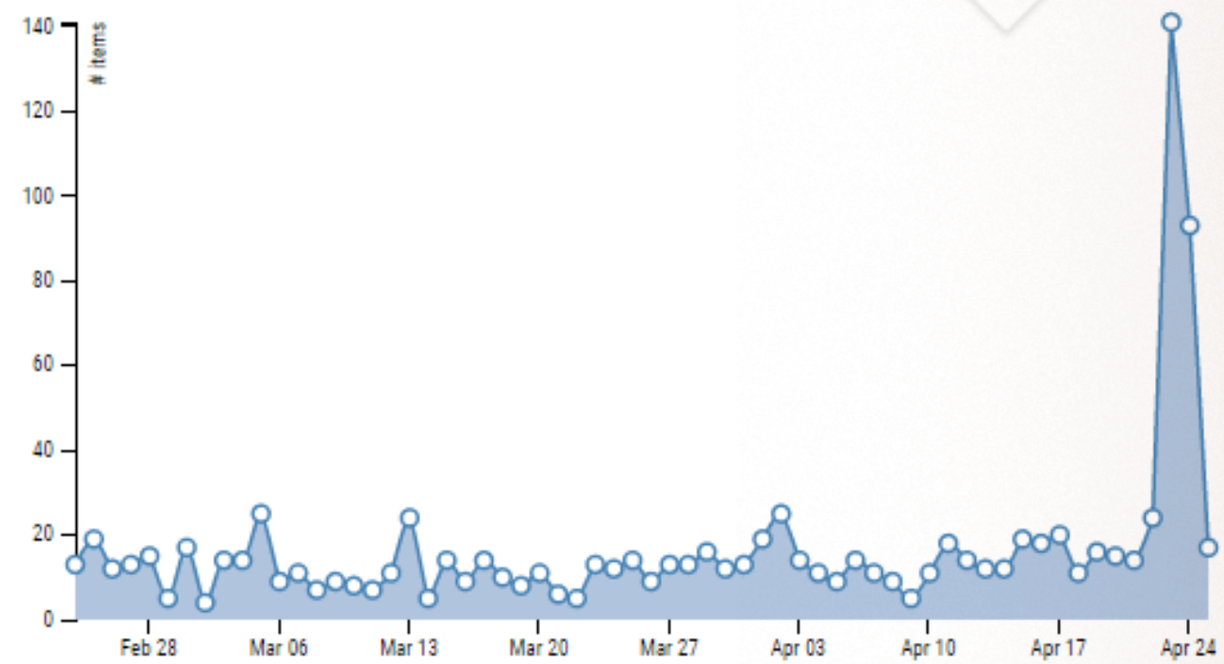
OUTPUT
1006 r

PREVIEW
3

	created_at	from_user_name	text
321458.0	2016-04-22T10:37:38	jafrater	#Shakespeare's grave at Holy Trinity in #Stratford #shakespeare400 @ Holy Trinity Church,... https://www.instagram.com/p/BEf9uSlrlUz/
595109.0	2016-04-25T14:38:29	MiguelWrites	We have indeed been celebrating

VISUALIZE RESULT

CREATED AT - DAILY VOLUME



ADD VIZ

INPUT
152610 r

PREVIEW
3

	created_at	from_user_name	text	filter
937.0	2016-04-24T08:18:57	Jeremiahjones88	LIVE on #Periscope: Sonnet Sunday #shakespeare's bday https://www.periscope.tv/w/aesncDFIV0t5TnpMTVZqQWd8MUJkR1IEVIZYem9LWG0EqIIRES2vMPIQg5HxWYX6we_6DVazsiFQ7Ov06IqY	
382.0	2016-04-12T1		RT @britishlibrary: Here's a sneak peek at our exhibition, opening this Friday. What's your favourite #Shakespeare	

Input preview

CODE YOUR FILTER

```
output = input.filter(function(item, i){

    if (
            // Check that latitude and longitude are valid
            item.lat && item.lng
            && !isNaN(+item.lat) && item.lat >= -90 && item.lat <
            && !isNaN(+item.lng) && item.lng >= -180 && item.lng
    ) {
        // If they are, concatenate them for the vis module
        item.coordinates = item.lat + '::' + item.lng
        return true
    } else {
        // If not, filter them
        return false
    }

});
```

Coding panel

OUTPUT
1006 r

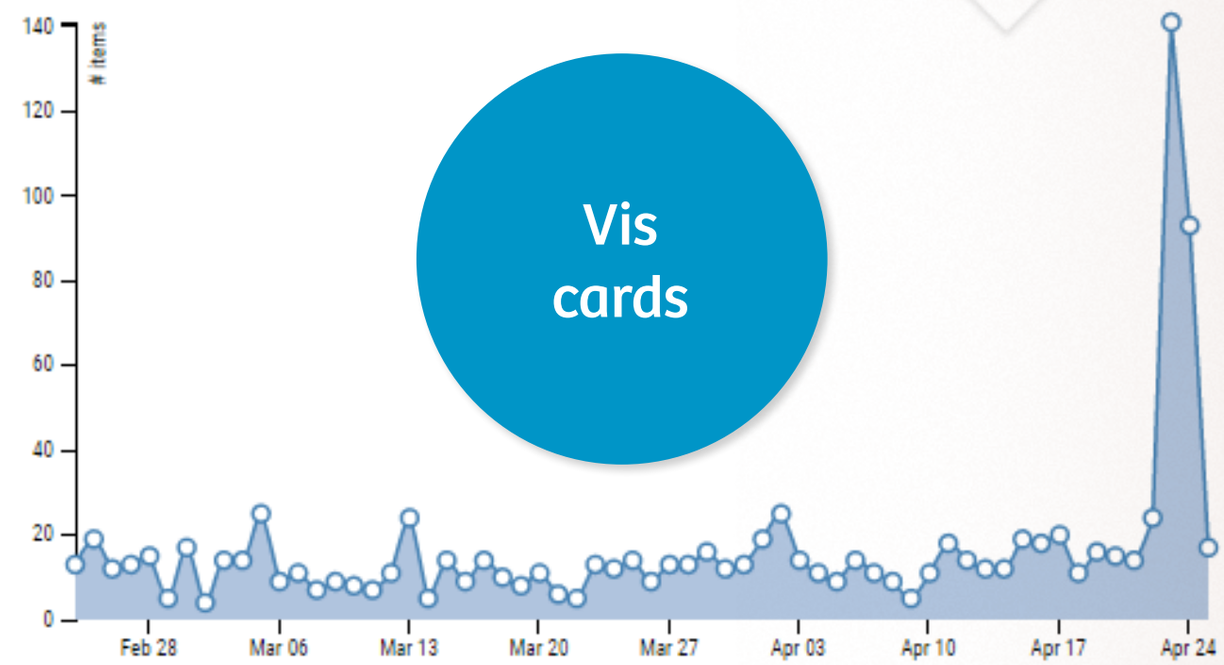
PREVIEW
3

	created_at	from_user_name	text
321458.0	2016-04-2		#Shakespeare's grave at Holy Trinity in #Stratford #shakespeare400 @ Holy Trinity Church,... https://www.instagram.com/p/BEf9uSIrlUz/
595109.0	2016-04-25T14:38:29	MiguelWrites	We have indeed been celebrating

Output preview

VISUALIZE RESULT

CREATED AT - DAILY VOLUME



Vis cards

+ ADD VIZ

INPUT
152610 r

PREVIEW
3

	created_at	from_user_name	text	filter
937.0	2016-04-24T08:18:57	Jeremiahjones88	LIVE on #Periscope: Sonnet Sunday #shakespeare's bday https://www.periscope.tv/w/aesncDFIV0t5TnpMTVZqQWd8MUJkR1IEVIZYem9LWG0EqIIRES2vMPIQg5HxWYX6we_6DVazsiFQ7Ov06IqY	
382.0	2016-04-12T16:26:22	Lynchypoos	RT @britishlibrary: Here's a sneak peek at our exhibition, opening this Friday. What's your favourite #Shakespeare	

CODE YOUR FILTER

```
output = input.filter(function(item, i){

    if (
          // Check that latitude and longitude are valid
          item.lat && item.lng
          && !isNaN(+item.lat) && item.lat >= -90 && item.lat <
          && !isNaN(+item.lng) && item.lng >= -180 && item.lng
    ) {
        // If they are, concatenate them for the vis module
        item.coordinates = item.lat + '::' + item.lng
        return true
    } else {
        // If not, filter them
        return false
    }

});
```

OUTPUT
1006 r

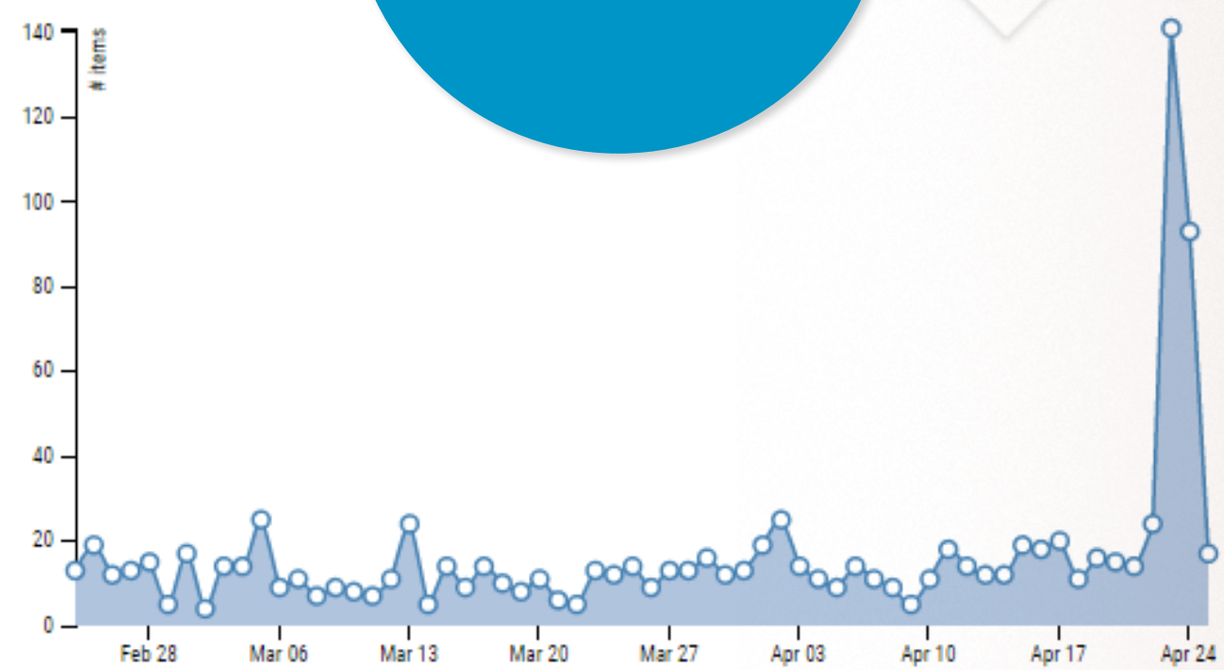
PREVIEW
3

	created_at	from_user_name	text
321458.0	2016-04-22T10:37:38	jafrater	#Shakespeare's grave at Holy Trinity in #Stratford #shakespeare400 @ Holy Trinity Church,... https://www.instagram.com/p/BEf9uSIrlUz/
595109.0	2		We have indeed been celebrating

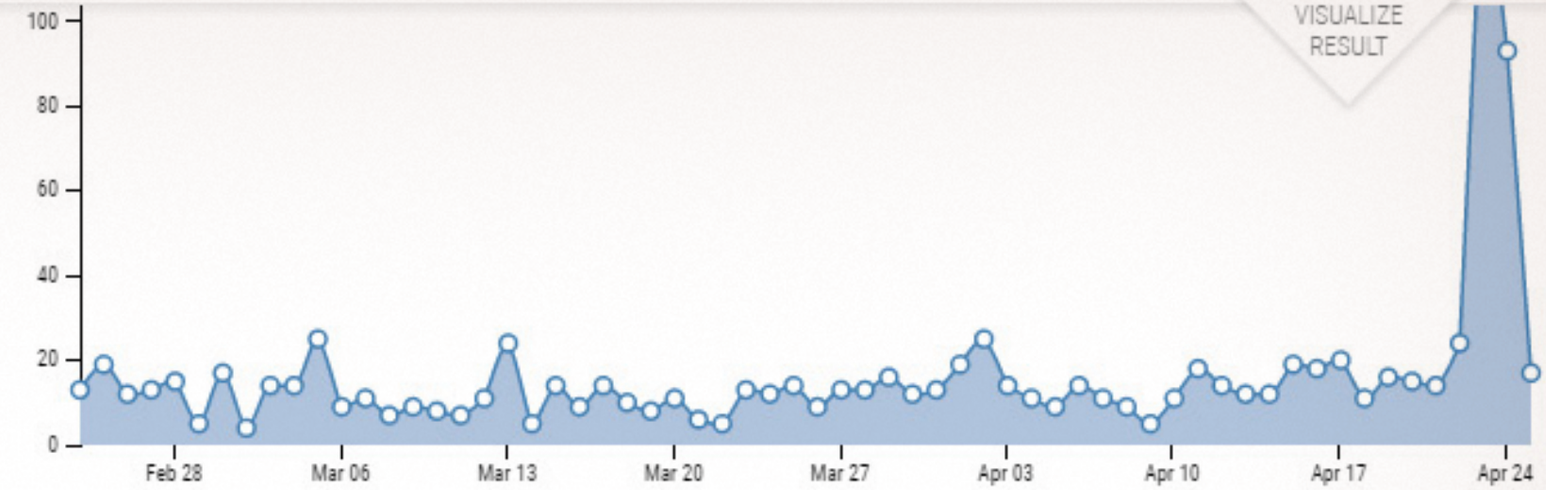
You can download the modified CSV

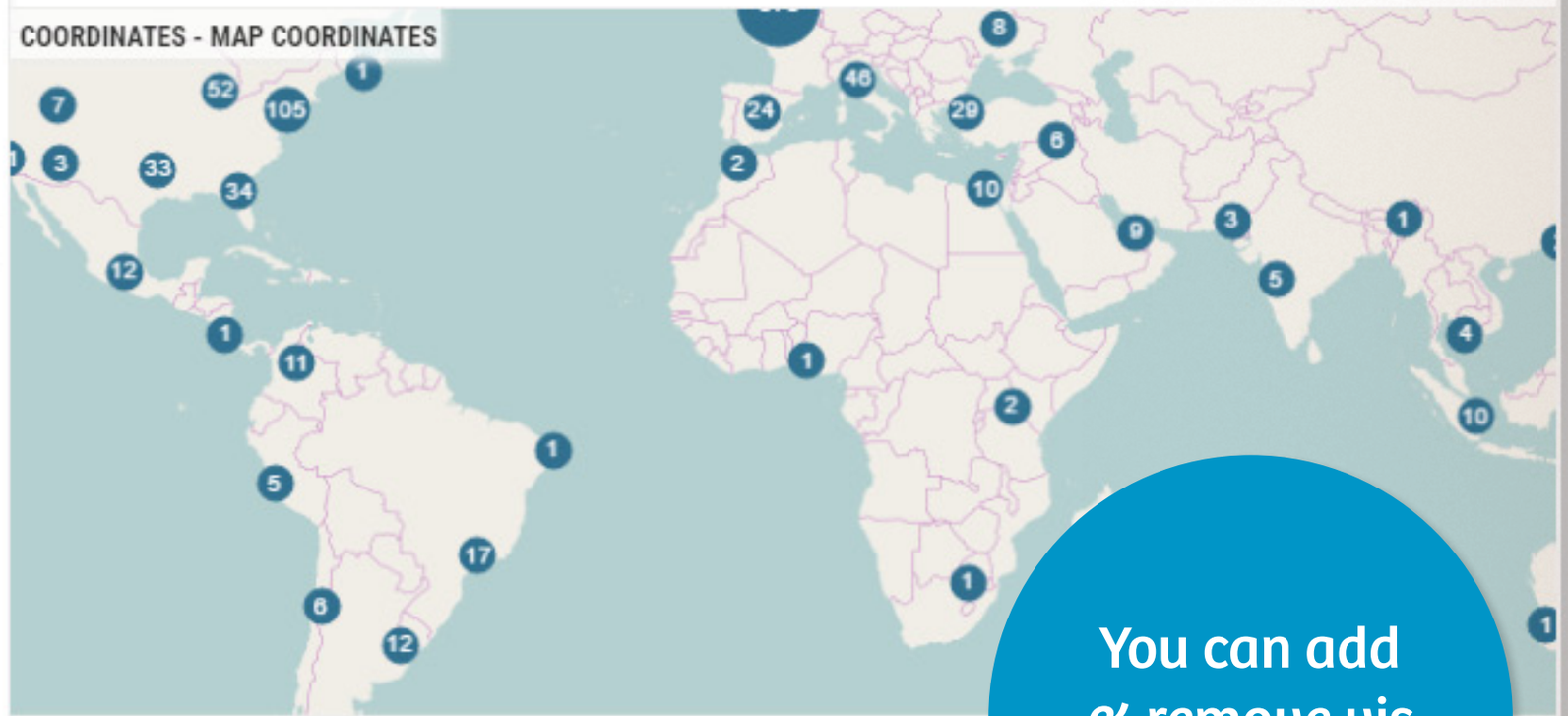
CREATED AT - DAILY VOLUME

VISUALIZE RESULT



ADD VIZ

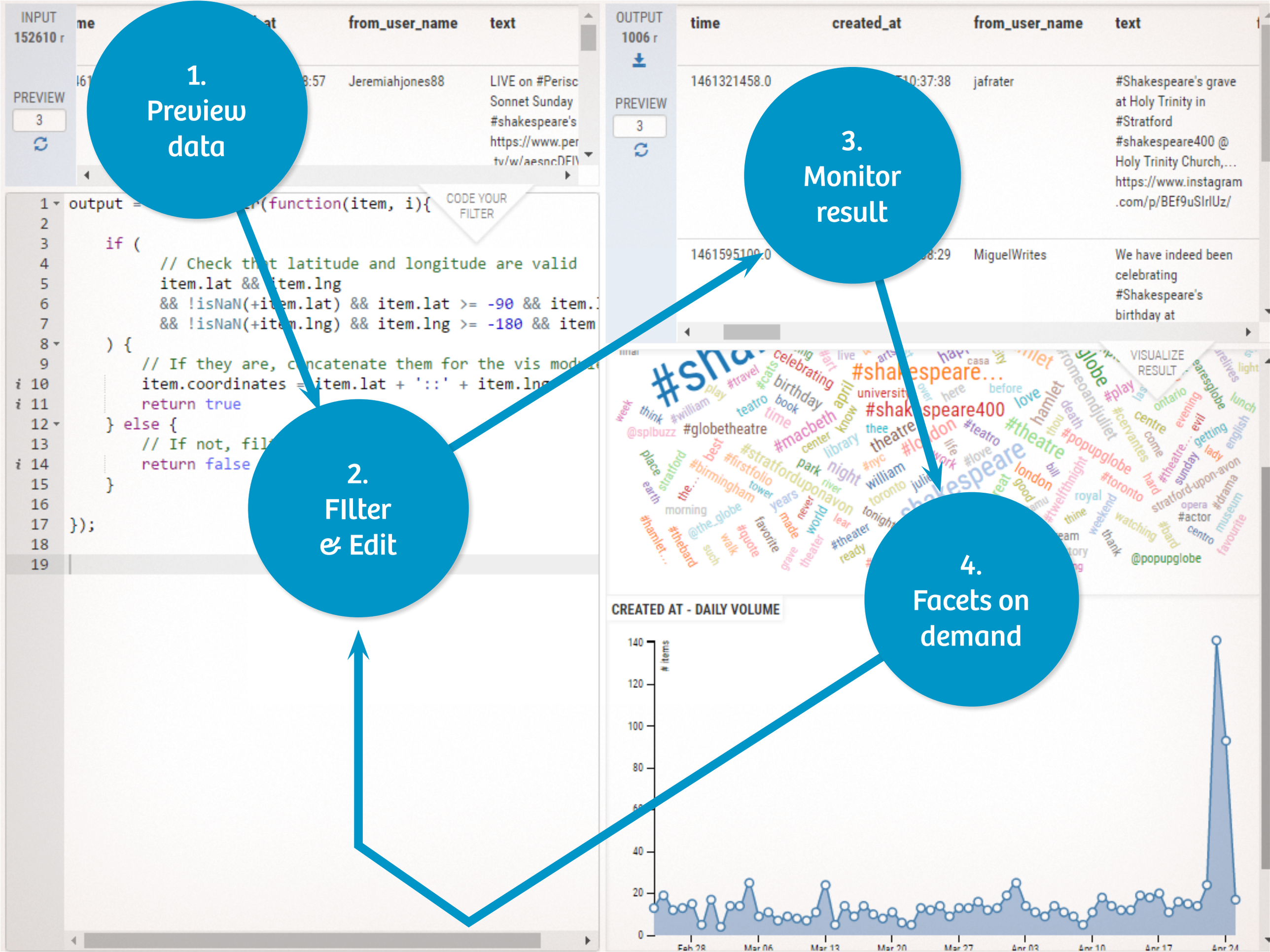


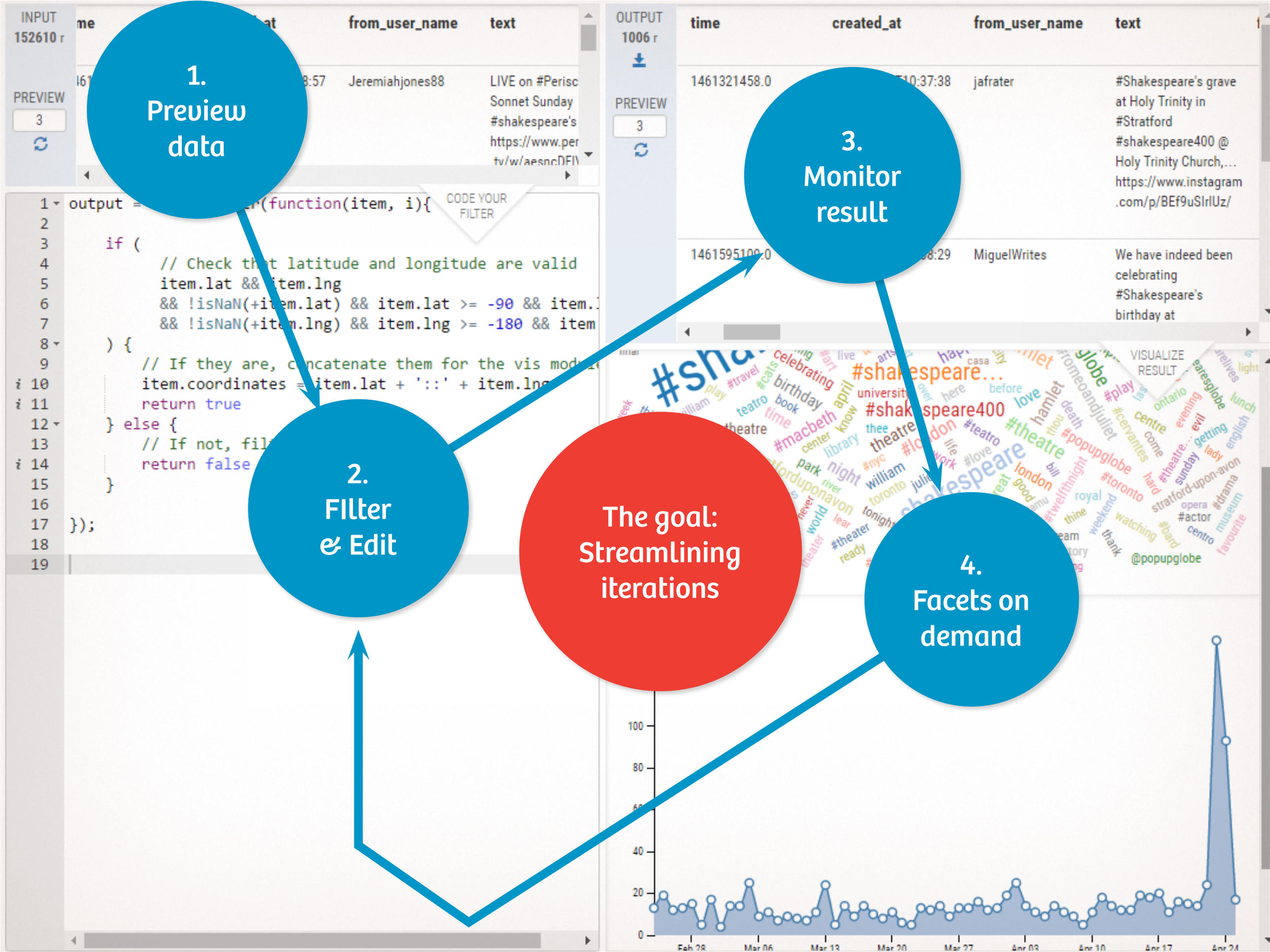


```
output = input.filter(funct , i){

    if (
        // Check that latitude and longitud
        item.lat && item.lng
        && !isNaN(+item.lat) && item.lat >=
        && !isNaN(+item.lng) && item.lng >=
    ) {
        // If they are, concatenate them for
        item.coordinates = item.lat + '::' +
        return true
    } else {
        // If not, filter them
        return false
    }

});

```

You can resize or hide the panels to fit your needs

You can add & remove vis cards


1. Preview data
2. FIlter & Edit
3. Monitor result
4. Facets on demand
INPUT
152610 r
PREVIEW
3
from_user_name
text
Jeremiahjones88
LIVE on #Perisc
Sonnet Sunday
#shakespeare's
https://www.per
CODE YOUR FILTER
output = r(function(item, i){
if (
// Check that latitude and longitude are valid
item.lat && item.lng
&& !isNaN(+item.lat) && item.lat >= -90 && item.
&& !isNaN(+item.lng) && item.lng >= -180 && item
) {
// If they are, concatenate them for the vis mod
item.coordinates = item.lat + '::' + item.lng
return true
} else {
// If not, fil
return false
}
});
OUTPUT
1006 r
PREVIEW
3
time
created_at
from_user_name
text
1461321458.0
10:37:38
jafrater
#Shakespeare's grave at Holy Trinity in #Stratford #shakespeare400 @ Holy Trinity Church,... https://www.instagram.com/p/BEf9uSIrlUz/
MiguelWrites
We have indeed been celebrating #Shakespeare's birthday at
VISUALIZE RESULT
CREATED AT - DAILY VOLUME
items
140
120
100
80
20
0
Feb 28
Mar 06
Mar 13
Mar 20
Mar 27
Apr 03
Apr 10
Apr 17
Apr 24


INPUT
152610 r
PREVIEW
3
me
at
from_user_name
text
Jeremiahjones88
LIVE on #Perisc
Sonnet Sunday
#shakespeare's
https://www.per
1. Preview data
CODE YOUR FILTER
output = r(function(item, i){
if (
// Check that latitude and longitude are valid
item.lat && item.lng
&& !isNaN(+item.lat) && item.lat >= -90 && item.
&& !isNaN(+item.lng) && item.lng >= -180 && item
) {
// If they are, concatenate them for the vis mod
item.coordinates = item.lat + '::' + item.lng
return true
} else {
// If not, fil
return false
}
});
2. FIlter & Edit
The goal: Streamlining iterations
OUTPUT
1006 r
PREVIEW
3
time
created_at
from_user_name
text
1461321458.0
10:37:38
jafrater
#Shakespeare's grave at Holy Trinity in #Stratford #shakespeare400 @ Holy Trinity Church,... https://www.instagram.com/p/BEf9uSIrlUz/
1461595100.0
8:29
MiguelWrites
We have indeed been celebrating #Shakespeare's birthday at
3. Monitor result
VISUALIZE RESULT
4. Facets on demand
100
80
40
20
0
Feb 28
Mar 06
Mar 13
Mar 20
Mar 27
Apr 03
Apr 10
Apr 17
Apr 24

Example:
Twitter data about Shakespeare

SciencesPo
MÉDIALAB

INPUT
152610 r

PREVIEW
3

	time	created_at	from_user_name	text
8726097702912	1461452106.0	2016-04-23T22:55:06	ReDewhurst	Can someone please reassure me that I'm not a bad reader for not liking (as far as I know) #Shakespeare 😐
5302054563841	1461422679.0	2016-04-23T14:44:39	victortisp	RT @400Cervantes: En 2016 celebramos la universalidad y genialidad de los autores del #Quijote y #Hamlet #400Cervantes #Shakespeare400 http://twitter.com/400Cervantes/status/722454409788137473/photo/1
6512314859520	1461406279.0	2016-04-23T10:11:19	PenguinIndia	RT @ a

```
// FILTER YOUR DATA HERE

// Just fill the "output" variable using "input"
output = input.filter(function(item, i){

    // EXAMPLE: return the first 100 items
    return i < 100;

});

// Hit CTRL + ENTER to run the code
```

Not much to see at upload

Edit your code then press
CTRL + ENTER

(works only from coding panel)

Example:
Twitter data about Shakespeare

```
1
2  output = input
3
4  // Hit CTRL + ENTER to run the code
```

Let's filter nothing and take a look at different facets

Example:
Twitter data about Shakespeare

SciencesPo
MÉDIALAB

CODE YOUR FILTER

```
1
2  output = input
3
4  // Hit CTRL + ENTER to run the code
```

OUTPUT
152610 r

PREVIEW
3

row #	id	time	created_at	from_user_na
128070	709048398062735361	1457885286.0	2016-03-13T16:08:06	JoeyFizzy

VISUALIZE RESULT

Which one?

1. en (76332)
2. fr (17222)
3. es (10363)
4. pt (8204)
5. it (8116)

ITEMS TOP 50
List of the 50 most occurrent items

WORDS CLOUD
Word cloud obtained from basic text mining

16. excited (8369)
17. info (8326)
18. stay (8222)
19. tuned (8136)
20. #designer (77

WORDS TOP 50
The 50 most occurrent words, obtained from basic text mining

CALENDAR
Daily count of items in a calendar view

DAILY VOLUME
Daily count of items as a curve

MONTHLY VOLUME
Monthly count of items as a curve

YEARLY VOLUME
Yearly count of items as a curve

MAP COORDINATES
Geographical distribution of items with dynamic clustering. Coordinates format: '0.0::0.0' (latitude - longitude)

Add visualization: which one?

Example:
Twitter data about Shakespeare

SciencesPo MÉDIALAB

CODE YOUR FILTER

```
1
2  output = input
3
4  // Hit CTRL + ENTER to run the code
```

We chose «Daily Volume». Data from which column?

OUTPUT 152610 r

PREVIEW 3

row #	id	time	created_at	from_user_na
128070	709048398062735361	1457885286.0	2016-03-13T16:08:06	JoeyFizzy

VISUALIZE RESULT

DAILY VOLUME: Which column?

id | time | created_at | from_user_name | text | filter_level | possibly_sensitive | withheld_copyright | withheld_scope | withheld_countries | truncated | retweet_count | favorite_count | lang | to_user_name | in_reply_to_status_id | source | location | lat | lng | from_user_id | from_user_realname | from_user_verified | from_user_description | from_user_url | from_user_profile_image_url | from_user_utcoffset | from_user_timezone | from_user_lang | from_user_tweetcount | from_user_followercount | from_user_friendcount | from_user_favourites_count | from_user_listed | from_user_withheld_scope | from_user_withheld_countries | from_user_created_at

Example:
Twitter data about Shakespeare

SciencesPo
MÉDIALAB

CODE YOUR FILTER

```
1
2  output = input
3
4  // Hit CTRL + ENTER to run the code
```

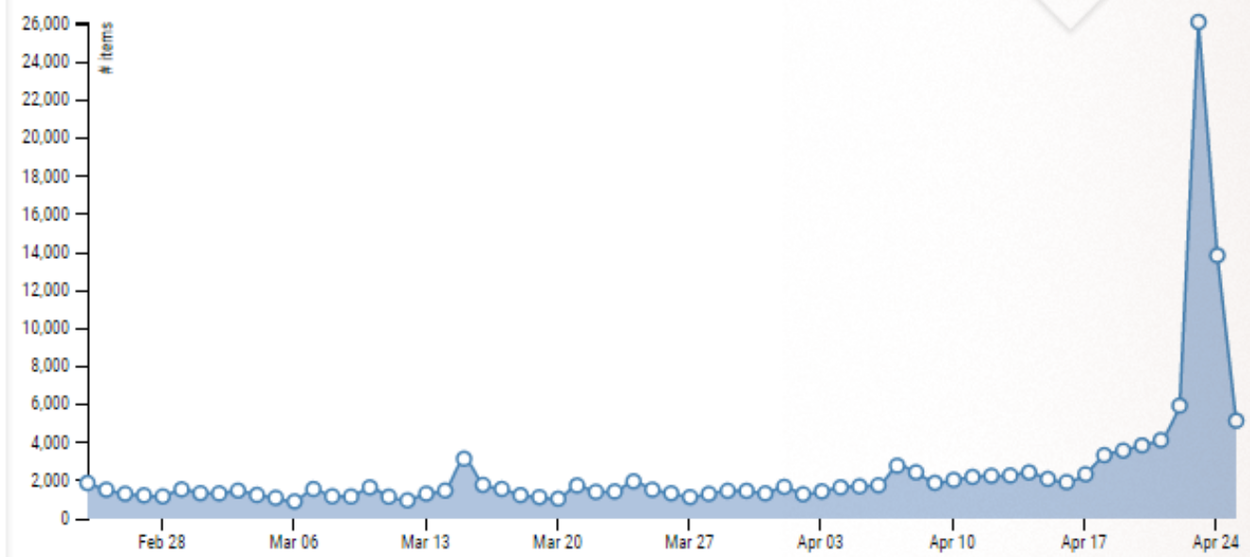
OUTPUT
152610 r

PREVIEW
3

row #	id	time	created_at	from_user_na
128070	709048398062735361	1457885286.0	2016-03-13T16:08:06	JoeyFizzy

VISUALIZE RESULT

CREATED AT - DAILY VOLUME



ADD VIZ

Column:
«created_at»
...the vis is
complete

Example:
Twitter data about Shakespeare

SciencesPo MÉDIALAB

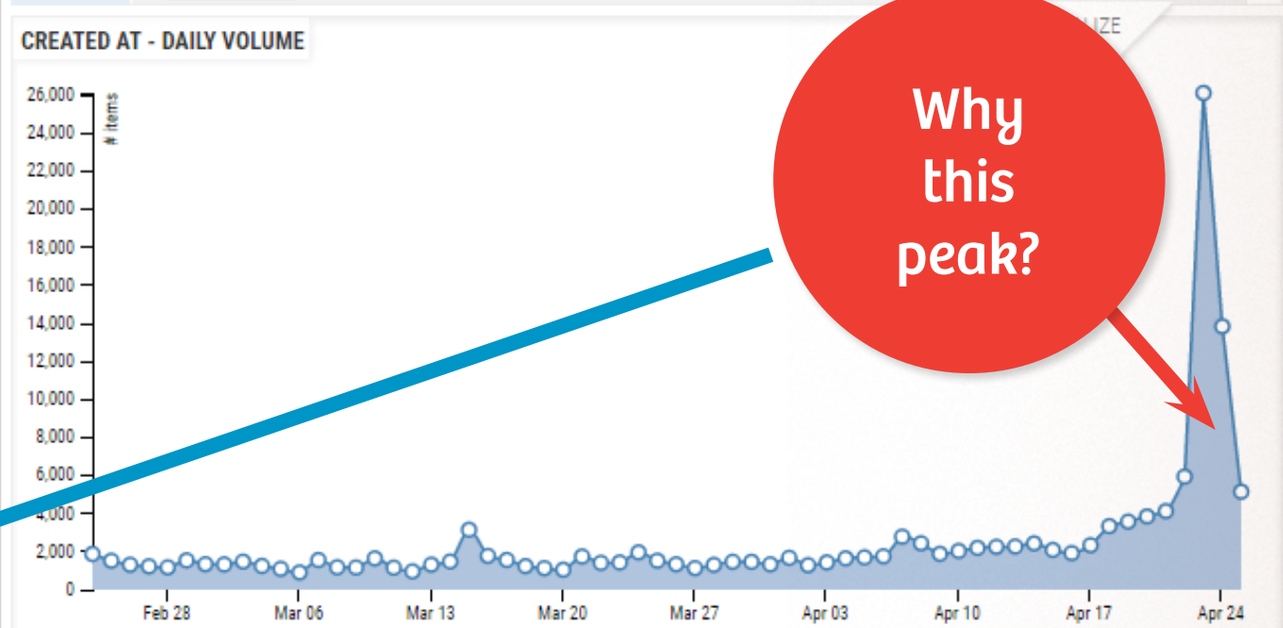
CODE YOUR FILTER

```
1
2  output = input
3
4  // Hit CTRL + ENTER to run the code
```

OUTPUT 152610 r

PREVIEW 3

row #	id	time	created_at	from_user_na
128070	709048398062735361	1457885286.0	2016-03-13T16:08:06	JoeyFizzy



Why this peak?

...Back to filtering

+ ADD VIZ

Example:
Twitter data about Shakespeare

```
var limitDate = new Date('2016-04-20')
output = input.filter(function(item, i){

    var date = new Date(item.created_at)
    return date > limitDate

});
```

We parse dates
and filter after the
2016-04-20

Example:
Twitter data about Shakespeare

SciencesPo
MÉDIALAB

CODE YOUR FILTER

```
var limitDate = new Date('2016-04-20')
output = input.filter(function(item, i){

    var date = new Date(item.created_at)
    return date > limitDate

});
```

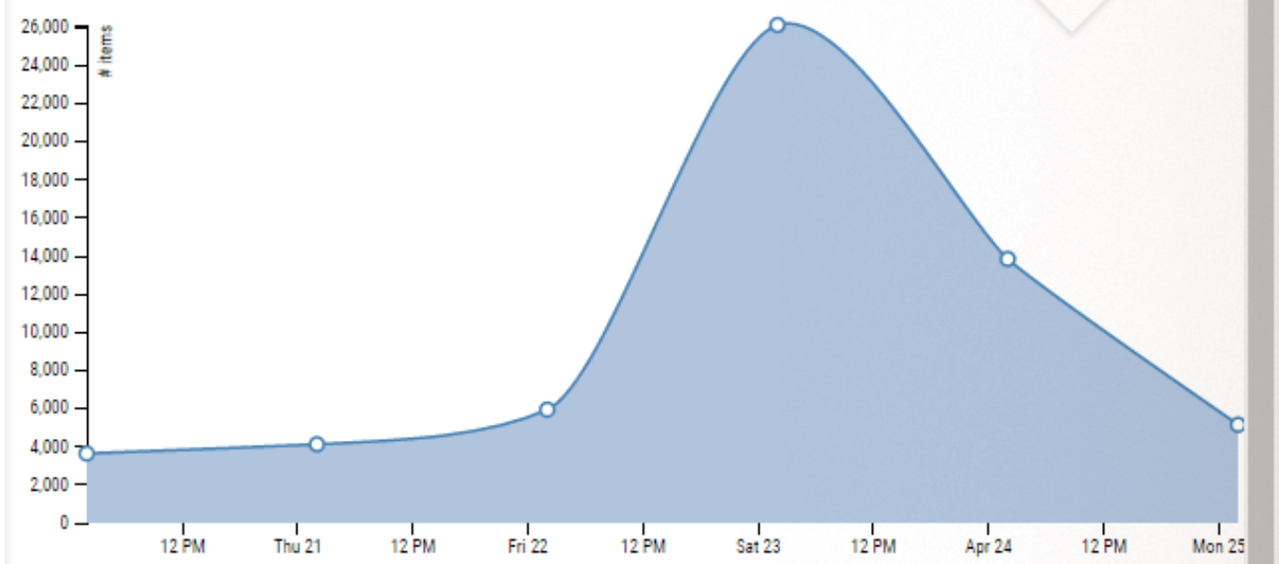
OUTPUT
58848 r

PREVIEW
3

	time	created_at	from_user_name	text	filt
797359617	1461536744.0	2016-04-24T22:25:44	peiweotck	RT @AkashaGarnier: "Boldness be my friend." ~ @Wwm_Shakespeare #Shakespeare400	

VISUALIZE RESULT

CREATED AT - DAILY VOLUME



TEXT - WORDS TOP 50

1. #shakespeare (51389)
2. #shakespeare400 (14623)
3. today (4496)
4. shakespeare (4124)
5. #hamlet (3824)
6. #shakespearelives (3548)
7. wise (3516)
8. fool (3515)
9. years (3428)
10. death (3206)
11. william (3086)
12. died (2928)
13. anniversary (2547)
14. #cervantes (2400)
15. 400th (2220)
16. think (2010)
17. celebrate (1934)
18. #macbeth (1912)
19. birthday (1898)
20. happy (1872)
21. himself (1821)
22. #rsclive (1801)
23. knows (1782)
24. doth (1781)
25. "the (1710)
26. here (1615)
27. merchant (1596)
28. venice (1591)
29. love (1535)
30. gold (1534)
31. glisters (1526)
32. @chakerkhazaal (1518)
33. more (1491)
34. plays (1417)
35. bard (1221)
36. @bbcle (1194)
37. celebrating (1147)
38. @dailyshakes (1143)
39. play (1079)
40. #onthisday (1054)
41. prince (1041)
42. still (1025)
43. phrases (969)
44. @thersc (946)
45. great (942)
46. course (907)
47. friend (860)
48. quotes (846)
49. learn (843)
50. read (842)

+ ADD VIZ

And we add a vis card to look at the content of the tweets

Example:
Twitter data about Shakespeare

SciencesPo MÉDIALAB

Before the peak

1. #shakespeare (84829)
2. shakespeare (10099)
3. #shakespeare400 (4721)
4. here (3376)
5. #hamlet (3141)
6. today (3128)
7. more (2902)
8. love (2679)
9. william (2588)
10. #cervantes (2489)
11. april (2469)
12. #macbeth (2190)
13. first (2058)

By iterating, we can compare

During the peak

1. #shakespeare (51389)
2. #shakespeare400 (14623)
3. today (4496)
4. shakespeare (4124)
5. #hamlet (3824)
6. #shakespearelives (3548)
7. wise (3516)
8. fool (3515)
9. years (3428)
10. death (3206)
11. william (3086)
12. died (2928)

Example:
Twitter data about Shakespeare

SciencesPo MÉDIALAB

Before the peak

1. #shakespeare (84829)
2. shakespeare (10099)
3. #shakespeare400 (4721)
4. here (3376)
5. #hamlet (3141)
6. today (3128)
7. more (2902)
8. love (2679)
9. william (2588)
10. #cervantes (2489)
11. april (2469)
12. #macbeth (2190)
13. first (2058)

During the peak

1. #shakespeare (51389)
2. #shakespeare400 (14623)
3. today (4496)
4. shakespeare (4124)
5. #hamlet (3824)
6. #shakespearelives (3548)
7. wise (3516)
8. fool (3515)
9. years (3428)
10. death (3206)
11. william (3086)
12. died (2928)

Its the anniversary of Shakespeare's death

Example:
Twitter data about Shakespeare

SciencesPo
MÉDIALAB

created_at	from_user_name	text
2016-04-23T18:45:45	ru_arena	RT @Libroantiguo: "The fool doth think he is wise, but the wise man knows himself to be a fool." #Shakespeare died #OnThisDay in 1616. http://twitter.com/Libroantiguo/status/723785774420865025/photo/1

The output monitoring helps validating the hypothesis

Wrap up

Exploring data requires iterating

That is why **CSV, Rinse, Repeat**
is about constantly rewriting filters

The visualizations are too basic, the preview is not comfortable...
That's fine! You don't really need more during exploration.
Our design aims at being «KISS»: Keep It Simple Stupid

Exploration is over when you have hypotheses.
At this point, just switch to a more analytical environment:
Libre Office, Tableau, R, Stata...

Thank you for your attention

Mathieu.Jacomy@sciencespo.fr

SciencesPo
MÉDIALAB

http://medialab.sciences-po.fr